

d·i·e

Deutsches Institut für
Entwicklungspolitik



German Development
Institute

Discussion Paper

8/2012

How to Evaluate Budget Support Conditionality and Policy Dialogue

Using the qualitative approach to causality

Ha Hoang

How to evaluate budget support conditionality and policy dialogue

Using the qualitative approach to causality

Ha Hoang

BMZ - DIE Research Project
“Development and Application of Evaluation Methods and
Approaches Project”

Bonn 2012



Discussion Paper / Deutsches Institut für Entwicklungspolitik
ISSN 1860-0441

Die deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.d-nb.de> abrufbar.

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data is available in the Internet at <http://dnb.d-nb.de>.

ISBN 978-3-88985-548-0

Ha Hoang was a researcher at DIE from 2009 to 2011. She holds Master of Art in Public Policy and in Development Evaluation and Management. Her areas of interest include public management, social policies, applied qualitative and quantitative research and evaluation methods.
E-mail: h.hoang@mpp.hertie-school.org

© Deutsches Institut für Entwicklungspolitik gGmbH
Tulpenfeld 6, 53113 Bonn
 +49 (0)228 94927-0
 +49 (0)228 94927-130
E-mail: die@die-gdi.de
<http://www.die-gdi.de>

Abstract

After decades of providing development aid, international aid agencies are still in search of more effective aid modalities. In this context, budget support has emerged as a potential candidate that could foster positive changes in poverty reduction. Unfortunately, the verdict on the effectiveness of this aid modality is still being discussed. By looking at budget support conditionality and policy dialogue, this paper argues that the application such qualitative methods to causality as Qualitative Comparative Analysis and Process-Tracing can go a long way towards answering the question of the effectiveness of budget support. When the causality chain is long and explanatory factors are numerous, as in the case of budget support, these methods can help to establish causality by tracing the causal pathways and causal mechanisms.

Ha Hoang

Bonn, March 2012

Contents

Abbreviations

1	Introduction: the need for a methodological approach to improve the evaluation of budget support policy dialogue and conditionality	1
2	History and intervention logic of budget support	3
2.1	History and context: budget support as a “new aid modality”	3
2.2	Intervention logic: the overall model and underlying hypothesis of how budget support is meant to work	5
2.3	Budget support conditionality, policy dialogue and qualitative methods of establishing causality	7
3	Evaluation frameworks	7
3.1	Evaluation framework for technical conditionality	7
3.2	Evaluating political conditionality and policy dialogue	37
4	Conclusions and recommendations	45
	Bibliography	49

Figures and Boxes

Figure 1:	Causal model of conditionality at implementation level	12
Figure 2:	Technical conditionality	24
Figure 3:	Measurement, design and analysis: pure and mixed combinations	30
Box 1:	Example of technical conditions	25
Box 2:	OECD DAC definition of the evaluation of effectiveness	25
Box 3:	Examples of the goal hierarchy of technical conditions	27
Box 4:	Example of governance effect profile for a technical condition	28
Box 5:	Process tracing: four tests for causation	35
Box 6:	Options for organizing and reporting qualitative data	42
Table 1:	Hypothetical truth table showing compliance with technical budget support conditionality	20

Abbreviations

BMZ	German Federal Ministry for Economic Cooperation and Development / Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung
BS	Budget Support
DFID	UK Department for International Development
EC	European Commission
GBS	General Budget Support
HIPC	Heavily Indebted Poor Country
IEG	Independent Evaluation Group
IMF	International Monetary Fund
MDBS	Multi-Donor Budget Support
OECD	Organisation for Economic Co-operation and Development
PAF	Performance Assessment Framework
PRSP	Poverty Reduction Strategy Paper
PD	Paris Declaration
PFM	Public Financial Management
PRSC	Poverty Reduction Support Credit
PBA	Programme-Based Approach
QCA	Qualitative Comparative Analysis
SPA	Strategic Partnership with Africa
UP	Underlying Principles
WB	World Bank

1 Introduction: the need for a methodological approach to improve the evaluation of budget support policy dialogue and conditionality

Budget support (BS) as an aid modality in its modern form of a “programme-based approach” (PBA) emerged in the late 1990s and early 2000s as a result of the international debate on how to make aid more effective. The financial transfers made by donors in the form of BS are channelled through the recipient government’s own budgetary system and are not earmarked for specific projects or expenditure items.¹ In addition to financial transfers, BS involves non-financial donor inputs, in particular policy dialogue, conditionality and technical assistance/capacity-building (de Kemp / Faust / Leiderer 2011).

As a financing instrument, BS is intended to support the partner country’s national poverty reduction strategy. BS policy dialogue centres on this broad-based reform agenda, budget allocations and policy implementation. By tying BS to performance as well as political conditionality, donors hope to provide incentives to implement pro-poor policies and governance reforms deemed necessary to improve overall government effectiveness. Concomitant technical assistance and capacity-building mainly target the national public financial management (PFM) system and external oversight bodies. The focus on partner countries’ PFM supposedly serves the dual purpose of reducing fiduciary risks inherent in providing budget support and building effective domestic institutions that are necessary for sustainable development processes. It is expected that, with this combination of measures in support of home-grown reforms, building local institutional capacity, providing fiscal and political space and offering adequate incentives to recipient governments, budget support will contribute to sustainable development and poverty reduction.

Past efforts to evaluate BS have so far produced only partially satisfactory insights into its effectiveness in achieving these objectives. In a first large-scale evaluation commissioned by 24 aid agencies and seven partner governments under the auspices of the Organisation for Economic Co-operation and Development with the Development Assistance Committee (OECD/DAC) evaluation network, an attempt was made to gauge “*to what extent and under what circumstances GBS is relevant, efficient and effective for achieving sustainable impacts on poverty reduction and growth*”, based on seven detailed case studies (Burkina Faso, Malawi, Mozambique, Nicaragua, Rwanda, Uganda and Vietnam). This evaluation found that BS had positive effects on harmonisation, alignment and policy development in all of the countries reviewed and positive effects on the allocative and operational efficiency of public expenditure and on public finance management (PFM) systems in all but two of the case studies (Dom 2007, 2). However, the evaluation team also found that it was unable to assess the ultimate poverty impact of BS, as it was not possible to “*confidently track distinct PGBS² effects to the poverty impact level in most countries*” (Dom 2007, 4). An evaluation of the World Bank’s BS instrument, the Poverty Reduction Support Credit (PRSC), conducted by the World Bank / IMF Independent Evaluation Group (IEG) came to a similar conclusion with regard to the poverty effects of BS, namely

1 In the form of sector budget support (SBS) BS is – often only notionally – earmarked for one or more specific sectors, but not to more specific expenditure items.

2 Partnership General Budget Support.

that it was not possible to identify PRSCs with growth outcomes and that attributing poverty reduction records to PRSC measures requires further exploration (IEG 2010, 73–74).

However, this IEG evaluation (IEG 2010) also leaves much to be desired with regard to its assessment of the effectiveness of the non-financial components of budget support, in particular conditionality and policy dialogue. While it does give a good description of implementation, revealing, for instance, whether conditionality accords with the national development strategy, the extent to which recipient governments have a sense of ownership for PRSC programmes, the level of harmonisation among BS donors, etc., it does not address the effectiveness question, and it remains unclear how the way the BS programme is delivered influences outcomes, how, for example, it affects conditionality compliance and the effectiveness of reforms or how it helps to shape government agendas.

In a similar vein, the joint OECD-DAC GBS evaluation attempted to follow the result chain, but failed to give a sufficiently concrete description of the crucial aspects of conditionality and dialogue to permit a clear link between conditionality compliance and changes in policy outcomes to be identified.

In recent years, donor governments have come under growing pressure to demonstrate results of their development cooperation at impact level. As a consequence, a number of donors have joined forces to make a fresh effort to evaluate BS even more comprehensively than in the first joint evaluation of 2005/2006. This new approach explicitly sought to cover the entire causal chain of BS from the various inputs (funding, policy dialogue, conditionality, technical assistance and capacity-building) to impact level (Caputo / Lawson / van der Linde 2008). It was piloted in three country studies (Mali, Tunisia, Zambia) between 2009 and 2011, with particular emphasis placed on evaluating the impact of sector policies backed by budget support, using a quantitative approach suggested by Gunning / Elbers / de Koop (2009). As a consequence, the authors of the three synthesis reports drawn up on these evaluations were much more confident that they had covered the entire causal chain from donor inputs to policies and policy impact and that positive changes in the form of growth and poverty reduction could be attributed to government policies partly funded by donors' budget support. At the same time, the studies are more or less frank about the persistent lack of a more robust methodology to evaluate the direct and indirect effects of the non-financial inputs of budget support, and particularly of BS policy dialogue and conditionality. In these studies, too, the attribution of causal links between these inputs and observed policy changes and reforms therefore had to be based on an informed judgement by the evaluators (see, for example, de Kemp / Faust / Leiderer 2011, 24).

This paper seeks to help to fill this methodological gap by examining the question of how to evaluate BS conditionality and policy dialogue more systematically. Specifically, it asks whether conditionality results in the implementation of reforms and the improvement of policy outcomes and the governance records of aid recipient countries and whether policy dialogue leads to positive changes in the partner government's political agenda. It considers at length what current methods are most likely to answer these evaluation questions.

The evaluation of such “soft inputs” as policy dialogue and conditionality faces many methodological challenges. For one thing, the nature of the intervention and the number of cases and observations usually make the evaluation task unsuitable for statistical and quantitative methods. Nor are there any satisfactory guidelines to show how qualitative methods can be used to assess the effectiveness of these inputs. What is more, there is a lack of systematic reviews of qualitative methods and how they can be employed to answer the evaluation questions concerned.

This paper was written after a systematic review of qualitative research and evaluation methods. Particularly valuable was the work of Michael Quinn Patton (2002) on qualitative methods. It provides excellent guidance on qualitative data analysis for descriptive purposes, which forms the “bedrock” of causal inference. The analysis in this paper also makes extensive use of more recent work by Goertz / Mahoney s. a.) and Andrew Bennet (2010), which advances process-tracing as a method by elaborating on the logic of causal inference.

The paper is divided into three parts. Chapter 2 describes the intervention logic of budget support in some detail. Chapter 3 constructs analytical frameworks for technical and political conditionality and for policy dialogue. An evaluation design follows each analytical framework, specifying case selection, data collection and analysis methods and ends with concluding remarks. Chapter 4 concludes and makes a number of recommendations.

2 History and intervention logic of budget support

2.1 History and context: budget support as a “new aid modality”

Budget support is defined as “*the transfer of financial resources of an external financing agency to the National Treasury of a partner country, following respect by the latter of agreed conditions for payment*” (Caputo / Lawson / van der Linde 2008). As such, BS in its modern form is not an entirely new aid modality, but has its roots in the structural adjustment programmes of the World Bank and International Monetary Fund (IMF) and policy-based lending operations (PLOs) of the World Bank, which partially replaced project-based aid in the period leading up to the 2000s. This latter shift occurred as a result of accumulated lessons on the effectiveness of project aid: among other things, project aid creates parallel systems that bypass and at times undermine the state apparatus, disbursement is low, and the development impact is fragmented and limited (Killick / Gunatilaka / Marr 1998; Koeberle / Stavreski 2006).

In its current form, BS can also be traced back to the World Bank’s and IMF’s Heavily Indebted Poor Country (HIPC) initiative, which drew on lessons learnt during the previous episode of structural adjustment programmes, when the Bretton Wood institutions provided direct budget support to help countries to weather balance-of-payment difficulties in exchange for structural reforms in the areas of macroeconomic management and market liberalisation. However, these programmes had little positive impact on growth, some

positive impact on export and balance-of-payment performance, but were also often associated with a negative impact on investment levels (Killick / Gunatilaka / Marr 1998). In addition, many programmes suffered from non-compliance with the conditions attached. Conflicts of interests between donors and recipient governments and deep-seated domestic institutional factors stood in the way of reforms. One lesson acknowledged today is that financial assistance cannot compensate for a lack of domestic willingness to reform.

Donors provide BS in its new form of poverty reduction budget support not in exchange for a set of policy actions as reforms, but for the implementation of partner countries' own national poverty reduction and development plans. Partnership has replaced the concept of "*top-down conditionality*" when it comes to the nature of the aid relationship. This new approach to conditionality is in line with the 2005 Paris Declaration (PD). The ultimate aim of the PD is to increase aid effectiveness and so to contribute to "*reducing poverty and inequality, increasing growth, building capacity and accelerating achievement of the MDGs*". The signatories to the PD uphold five principles for effective aid: ownership, harmonisation, alignment, managing for results and mutual accountability. As a programme-based approach (PBA) of the type envisaged by the PD, budget support became one of the vehicles for the implementation of the PD and its principles.

The five PD principles have a number of implications for the operational design of budget support and other PBAs, one being closer collaboration between budget support donors so that they may act within a common framework when it comes to planning, funding, disbursement, monitoring and evaluation, the aim being to reduce the transaction costs to be borne by the government (the principle of harmonisation). The remaining principles delineate the role of each side in ensuring effective aid management. For example, donors should maximise their utilisation of local systems of public financial management, procurement, result frameworks and monitoring, while partner countries should be responsible for providing accountable and transparent systems (the principle of alignment). The principles of mutual accountability and managing for results require partner countries to develop national development strategies that are linked to budgetary processes and frameworks for monitoring and assessing progress in a country-wide, participatory fashion. Donors, on the other hand, should link their programmes and resources to this result-oriented framework and provide reliable information on aid flows to facilitate partner governments' budgetary processes. In summary, budget support is no longer a simple exchange of financial support for policy reforms. It has instead become a concerted effort by donors to engage in a long-term process of institution-building that not only generates better public goods and services with the aim of reducing poverty, but also supports the establishment of an accountable governance system.

The following section investigates the underlying hypothesis and intervention logic of budget support, before turning to the essential question: how can we better evaluate the effectiveness of conditionality and policy dialogue?

2.2 Intervention logic: the overall model and underlying hypothesis of how budget support is meant to work

Caputo / Lawsen / van der Linde (2008) outline a comprehensive evaluation framework for budget support at six levels: context, inputs, directed outputs, induced outputs, outcomes and impacts. At the input level, BS donors transfer financial resources directly to the partner government's national treasury. In principle, these funds serve the purpose of helping the country to implement its poverty reduction strategy. The transfer of funds is, however, contingent on a number of conditions. If the ownership of reforms and the development agenda are to be respected, conditionality content should be derived as far as possible from the national development strategy or otherwise agreed with the recipient government. Such negotiations take place through policy dialogue, which focuses on the broad-based reform agenda, budget allocation and policy priorities. Conditionality should be so linked to financial assistance that the predictability of funds is increased and, at the same time, incentives to implement particular reforms and policies are created. This conditionality is now primarily *ex post* rather than *ex ante*, i.e. disbursements depend on achievement rather than government promises.

Technical assistance and capacity-building primarily target the national public financial management (PFM) system. A stronger PFM system will boost the effectiveness of the BS financial component by ensuring more efficient and effective use of public resources. A better framework for public policy and public expenditure will also cushion BS funds against fiduciary risks, since donors now rely on the often poor-quality PFM of partner countries for aid delivery, and it is not possible to identify the end use of funds (Shand 2006).

The principles of alignment and harmonisation govern the way BS programmes are executed. Alignment refers to the practice of deriving conditionality content from national development plans, providing financial support for such plans and utilising national systems in financial reporting, monitoring and procurement. The alignment of donors with a single national development strategy and budgetary process precludes the risk of the government system being bypassed and so helps to create ownership of reform programmes. It also facilitates budgetary coherence and comprehensive planning by making the national budgetary process a centrepiece in the delivery of public goods and services (BMZ 2008).

The principle of harmonisation requires donors to collaborate on content and adopt common arrangements at all stages of aid management, e.g. diagnostic reviews, planning, disbursement, monitoring and evaluation, and reporting. Better coordination is expected to reduce transaction costs for both donors and partner countries. In summary, the non-financial components contribute to the effectiveness of the financial component by making aid more predictable, reducing transaction costs, reserving the driving seat (ownership) for the partner country and, at the same time, building its institutional capacity in key areas (de Kemp / Faust / Leiderer 2011).

Beyond the input level: from direct outputs to results

The next two levels of the causal chain in the logical framework are direct outputs and induced outputs. Direct outputs indicate how BS components are delivered, e.g. increased size and share of funds, more coordinated and nationally aligned technical assistance, policy dialogue and conditionality. Activities at input level, through interaction with local actors and the budget and policy-making process, define the shape of BS programmes. At this level it can be determined whether BS funds are predictable, conditionality aligns with the country's development agenda, technical assistance is strategic, and policy dialogue focuses on the government's priorities.

The third level – induced outputs – refers to the changes that should occur at both sectoral and national level once a BS programme is implemented. These changes consist in the improvement of the financing and institutional framework for public spending and policy and thus of policy management and service delivery. The effects of BS are considered significant at this level (Compernelle / de Kemp 2009), since the interaction is direct and the causal chain is relatively short.

The next two levels are outcome and impact. The government's policies and strategies (to which BS contributes) interact with the wider economy to produce outcomes, while impacts are felt in the longer term in growth, poverty and other aspects of development. As the government is but one source of influence, it is difficult to isolate its contribution, let alone the contribution made by BS, to outcomes and impacts.

As for the effects of BS, some authors distinguish between endogenous and exogenous effects (ibid). Endogenous effects originate from the internal process of accountability. BS triggers this effect by highlighting the role of public financial management, i.e. through direct fund transfer and utilisation of the national system. Budget and policy priorities gain in importance and receive due attention from the parliament, civil society and the public. The finance ministry will play a greater role in coordination, while the line ministries will focus on better policy-making and consensus-building to secure the resources allocated to them. The exogenous effects of budget support result from the intervention of such non-financial inputs as policy dialogues, conditionality and technical assistance/capacity-building.

In general, the theory of budget support paints an ambitious picture of its effectiveness, but provides little detail on the underlying mechanism. The effects of the non-financial components are no longer visible after the first level of direct outputs, and it is not possible to separate the endogenous and exogenous effects of budget support after the induced output level. The current theory falls short of distinguishing between different actors, even within the government system, at the various stages of the policy processes, the nature and pattern of interaction among constellations of actors and their consequences for policy outcomes. Very few attempts have been made to shed light on the underlying interaction processes that may well cause budget support to have the effects that it has. Much research needs to be done to unearth the mechanisms at work in this long causality chain from input

to impact, so that a convincing and realistic picture may be produced of how budget support can and in fact does work.

2.3 Budget support conditionality, policy dialogue and qualitative methods of establishing causality

This paper considers how to evaluate the effectiveness of budget support conditionality and policy dialogue using qualitative methods to establish causality. Conditionality and policy dialogue contribute directly to the effectiveness of budget support. However, most evaluations have so far concentrated on the effects of the financial components, and much less effort has been expended on the evaluation of conditionality and policy dialogue, despite their crucial role in the intervention logic for this aid modality. In conditionality, the concept of partnership replaces that of top-down conditionality, and the links between compliance and fund transfer are designed to make aid more predictable and create appropriate incentives. As the theory goes, stronger ownership and more predictable aid will boost the effectiveness of the financial component in reducing poverty. Policy dialogue, on the other hand, provides unprecedented opportunities for donors to engage with partner governments on macro and strategic policy and budgetary issues.

The choice of qualitative methods to establish causality has its roots in several observations. Conditionality and policy dialogue produce data and information that cannot be easily or meaningfully quantified. Conditionality and policy dialogues are characterised by complex political processes, with many actors and factors interacting. As such, qualitative methods include tools for dealing with this complexity. For example, qualitative comparative analysis can detect different causal pathways to similar outcomes, and process-tracing can pinpoint causes and effects by identifying causal mechanisms. As qualitative methods offer insights into causal mechanisms and causal processes, they can disclose unintended effects, uncover important explanatory factors and explain why things have or have not happened. These are the unique advantages of the method, and they are able to identify salient policy lessons fresh from the field. The next chapter will turn to the question of how to evaluate the effectiveness of conditionality.

3 Evaluation frameworks

3.1 Evaluation framework for technical conditionality

To begin with, a few conceptual clarifications are needed. The seemingly most striking paradox in evaluating budget support conditionality lies in its contrast with “*ownership*” – an overriding theme of this aid modality. Why should we still discuss, and even evaluate, the “*effectiveness*” of conditionality when this very notion has been replaced by “*partnership*” and “*ownership*”? A number of authors (rightly) argue that budget support “*increases*”, or at least remains as strong, rather than toning down or phasing out conditionality. For example, Hayman notes that the current generation of conditionality

hinges on “*a more complex set of norms and the underlying concepts of conditionality change very little*” (Hayman 2011, 679). The Poverty Reduction Strategy Paper is de facto a mega-condition for the participation of low-income countries in the HIPC initiative and has become one condition for the engagement of donors in budget support. In addition, attached to budget support are both the structural-adjustment-type conditions and those of the second wave of conditionality, i.e. political governance conditions. The difference is that donors should not impose these conditions from outside, but derive them from the partner countries’ policy documents (Hayman 2011).

If we examine the concept of conditionality in depth, we find that, while budget support donors may relax different parts of the concept to varying degrees, they embrace the notion as a whole. At basic level, conditionality comprises three components relating to the financing, the substance and the contingency of the former on the latter. Hard-core conditionality is one empirical example of the concept: “*policy changes stipulated as a prerequisite to the approval of, or continued access to, a grant or loan, or to subsequent assistance ... the expectation must be that the borrowing government would not voluntarily undertake the changes required*” (Killick 1997, 478). Finance plays a key role in leveraging policy actions here.. Yet donors differ in their interpretation of conditionality (Faust / Koch / Leiderer 2011). Some budget support donors, such as the UK, have temporarily shied away from this type of relationship and relaxed the contingency between financing and policy actions to the status of “*backing-up*” (DFID 2005). Other donors, Germany among them, have maintained a strong position on the relationship between financing and substance: “*conditions act as an incentive for the continuation of political reforms ... they fulfil a signal and control function by making the reform process and its successes more transparent and facilitating impact monitoring*” (BMZ 2008, 18).

In short, while budget support donors may tamper with different parts of the conditionality concept, they remain faithful to its overall structure: the contingency of financing on substance must remain in one form or another. Besides, as will be shown below, ownership and conditionality are overlapping concepts. As such, it is relevant to evaluate the effectiveness of budget support conditionality. This aim is also particularly relevant where there is no evidence of conditionality having enhanced ownership or programme success in the previous generation of programme aid (Dreher 2009) and the “new” generation of conditionality sets out to correct past mistakes.

As the Poverty Reduction Strategy Paper can be rightly categorised as part of the substance condition attached to budget support, this paper will focus solely on technical and political conditions. To assess the effectiveness of conditionality in the implementation of Poverty Reduction Strategy Paper (PRSPs), the whole budget support model would have to be evaluated, which is beyond the scope of this paper. Moreover, technical and political conditions are closely associated with the disbursement of BS financial resources and qualify as an essential aspect of conditionality.

Technical conditionality

Technical conditionality refers to the practice of making BS funds contingent on a set of specific policy-related conditions. These conditions form the content of the Performance Assessment Framework (PAF) – a mutually agreed policy matrix. The disbursement by BS donors of some of their funds is contingent on the degree to which partner governments comply with this policy matrix. Technical conditions typically fall into two categories: policy-based and result-based. Policy-based conditions denote the policy measures and actions that the partner government must take, such as passing legislation or restructuring a ministry. Result-based conditions are such policy outputs and outcomes as enrolment rates, numbers of health workers in clinics and health centres and vaccination/immunisation rates. Policy-based conditions stand at the input level on the policy process ladder: *“input indicators measure the actions and financial, administrative, and regulatory resources which are put into the development process”* (Schmidt 2006, 70). In result-based conditions, the output indicators measure *“the concrete and immediate consequences of the measure taken and resources used”*, while the outcome indicators measure *“the results and positive changes at the target-group level”* (ibid). The World Bank continues to be the strongest proponent of the former, while the European Commission is seen as the most vehement advocate of the latter.

Another dimension of the technical conditions is their coverage of policy fields. In this regard, we see an expansion from macroeconomic management to virtually all policy fields under the umbrella of poverty reduction. For example, the World Bank’s conditionality in 87 operations from 2001 to 2008 is spread over ten policy fields, among them agriculture, education, health, industry and trade and transport (IEG 2009).

As for the contingency of financing on the substance condition, different disbursement mechanisms exist. One common feature stands out: rewarding is now ex post rather than ex ante, i.e. financial assistance is disbursed according to performance achievement, not at the time promises are made. The European Commission (EC) operates one type of disbursement mechanism: funds are released as fixed tranches and variable tranches. A fixed tranche represents the all-or-nothing approach and is based on an assessment of macroeconomic stability and public financial management under the IMF’s Poverty Reduction and Growth Facility programme. The variable tranche is linked to the performance of specific indicators in PFM and social service delivery (technical conditions in PAF). The World Bank disburses its funds in a single tranche based on policy actions that have already been taken. Those actions are drawn from the list of triggers derived from the previous operation, which are not legally binding and can be revised, adapted, modified, replaced or dropped. Germany and a number of other bilateral donors operate the third type of disbursement mechanism, whereby donors release funds from the fixed tranche on the basis of an assessment of both technical and political conditions. Donors using the fourth type add an “incentive tranche”, which may be disbursed when the achievement level of PAF indicators exceeds 80 per cent.

The need for an evaluation framework

In the present context, evaluation framework means a set of hypotheses and evaluation questions, followed by a research and evaluation design aimed at testing the hypotheses and answering the evaluation questions. Such a framework for evaluating the effectiveness of budget support conditionality and policy dialogue do not yet exist. In the latest version of the budget support logical framework (Caputo / Lawsen / van der Linde 2008), the effects of conditionality merge with those of other BS and government inputs from the induced output level onwards. Such invisibility of the effects of segregated inputs can be tolerated at generic level, on the ground that *“budget support is not a development programme per se, but an aid modality that supports the development strategy of the beneficiary government”* (Caputo / Lawsen / van der Linde 2008, 11). However, if we are interested in evaluating the effectiveness of conditionality, we need to know in concrete terms to what exactly conditionality seeks to contribute, by what means and using what mechanism.

The question of the effectiveness of BS conditionality calls for causal thinking and a causal approach to evaluation, like any other form of impact/outcome evaluation. In other words, conditionality has its own intervening hypotheses, i.e. its objectives and how it intends to achieve them. It is necessary, therefore, to develop an ex-ante understanding of the hypothesised causes and effects and make explicit the underlying assumptions, which will be confirmed, refuted or revised as the result of the evaluation process.

The challenge of building such an analytical framework for BS conditionality stems from the fact that the task cannot be comfortably assigned to research or evaluation. Programme evaluation is *“the systematic collection of information about the activities, characteristics, and outcomes of programs to make judgements about the program, improve program effectiveness, and/or inform decisions about future programming”* (Patton 2002, 10). While such a task fits nicely with the evaluation of the effectiveness of BS conditionality, the intervention in and of itself is not specific enough to qualify for the task of programme evaluation. In other words, the absence of a coherent theory about how resources, activities and inputs transform through processes into intended outcomes prevents BS conditionality from being readily subjected to programme evaluation. Beyond the implementation level, it is not clear how different components of conditionality work with such other cross-cutting themes as alignment and harmonisation to produce effects at subsequent levels, e.g. *“improvements in policy processes, including the quality of policies and policy implementation”* (Caputo / Lawsen / van der Linde 2008, 20).

While the objective of generating a new theory is not the objective of the overall task, some aspects of the research approach are useful. First, when programme documents do not provide enough guidance, we should turn to research findings to clarify the causes and effects. Vague and sometimes contradictory concepts of causes and effects not only confuse evaluators as to what to measure, but also discredit the evaluation work, since they pose serious threats to the validity of the causal conclusion. Second, the research approach adopted in the qualitative method offers powerful tools for identifying causes and clarifying the causal mechanisms.

Implementation level

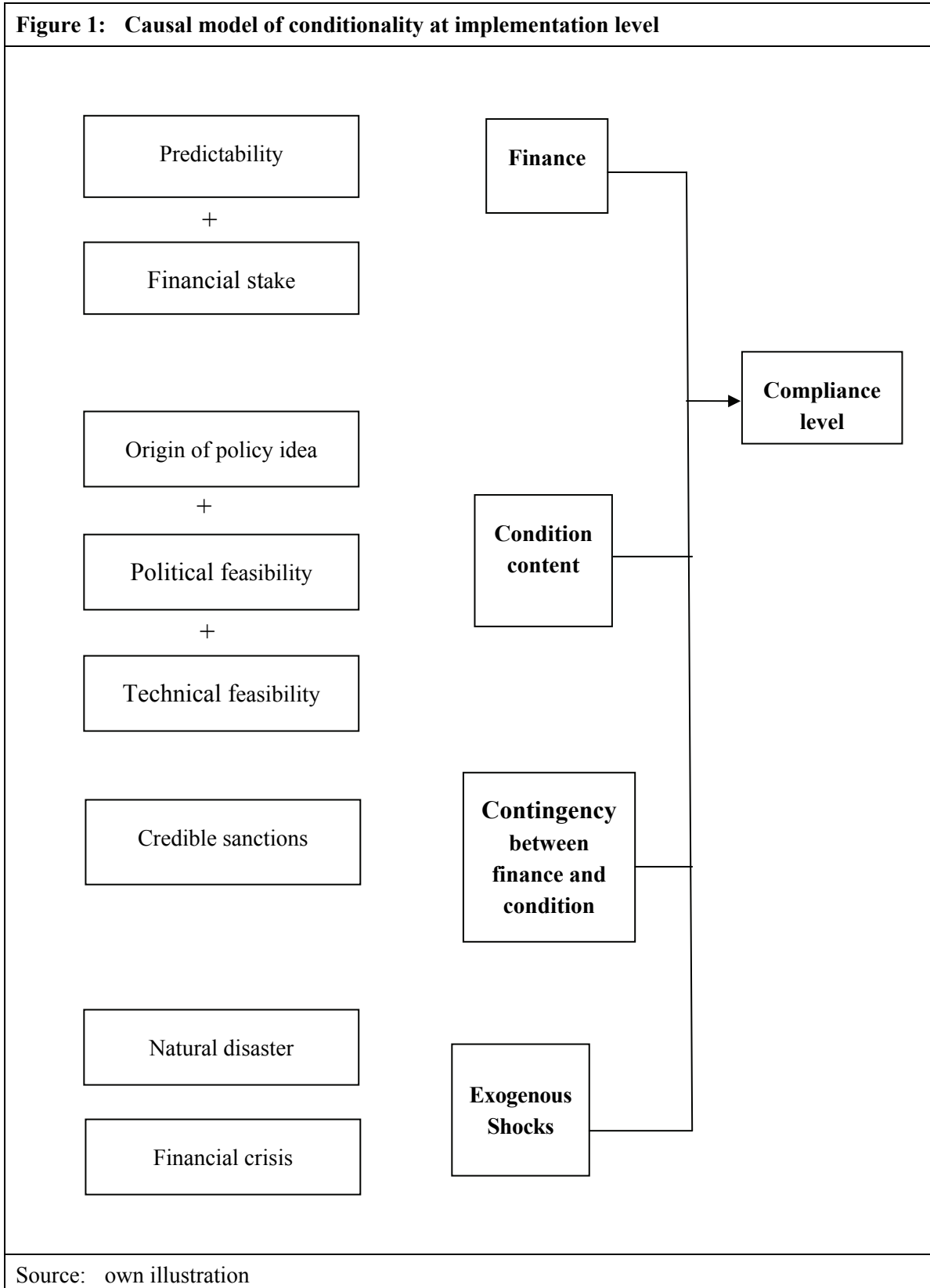
Before the question of programme effectiveness is considered, it must be asked how the programme has been implemented. Implementation refers to the actual operation of a programme, and the evaluation of implementation provides an opportunity for scrutinising how the programme has been delivered, especially in relation to the theory associated with it. At this stage, an evaluator can also study how organisational and external factors influence the programme, its activities and the achievement of its outcomes (Love 2004, 63–66). Implementation evaluation offers insights into the programme delivery process and thus shows how the programme can be improved, an important task of which the “black-box” paradigm of experimental evaluation is incapable (ibid).

Several methods of implementation evaluation have been employed in the case of budget support. For example, the Strategic Partnership with Africa (SPA) surveys regularly include *performance monitoring*, in which information is collected on certain aspects of budget support operations, e.g. alignment. Dijkstra’s work on PRSPs, ownership and aid effectiveness (Dijkstra 2011) gives an excellent example of the combination of *process evaluation* and *descriptive evaluation*. The author compares cross-sites (Bolivia, Honduras and Nicaragua) and reveals serious discrepancies between the planning and theory of PRSP formulation and actual delivery. From this evaluation, Dijkstra infers a fundamental defect in the PRSP approach: its orthodox planning methodology subordinates politics and depoliticises the development debate – a view echoed by other authors (Faust 2010). As a result, “donors operated in a self-created virtual, or pseudo, reality” parallel with the real local political discourse.

In BS conditionality, implementation evaluation entails asking (i) how conditionality has been delivered, (ii) what is the compliance level and (iii) what factors account for success or failure in respect of compliance. The first question will check whether the actual delivery of conditionality matches its intended plan and purposes, whether, for example, the content of the conditions reflects partner countries’ priorities, funds are linked to conditionality compliance in a predictable manner, etc. The second question assesses the compliance level of conditionality, i.e. how successfully partner governments achieve policy outcomes and reforms. The third question concerns attribution, i.e. it investigates and identifies factors that facilitate or inhibit the compliance level. The answers to this question are a source of policy learning about how to improve future programmes. Taken together, the three questions will tell a causal story about conditionality at implementation level.

Dependent variable: how conditionality has been delivered

The construction of the conditionality concept follows Goertz’s guidance on structuring concepts. Goertz’s approach draws attention to the important properties of the concepts that have causal power and are “*relevant to hypotheses, explanation and causal mechanisms*” (Goertz 2006, 4).



The secondary level of a concept consists of attributes to which causal power is often assigned. At this level, conditionality has three components: the funding or financial aspect, the conditions attached to the funding and the contingency of the two components. Goertz defines two types of concept structure: necessary and sufficient versus family resemblance. The former is non-substitutable, the latter substitutable. In the necessary and sufficient structure, for example, the absence of one dimension will exclude the case from the concept (Goertz 2006, 37). The concept of conditionality therefore has the necessary and sufficient structure. Each of the three components is necessary, and together they are suffice to form the concept. Cases in which BS donors provide funds without any conditions attached, prescribe desirable policy changes without any finance or do not link finance and conditions do not therefore count as conditionality.

At the third level of the concept structure is the indicator level. While the basic and secondary level indicates what is important about a concept, it often remains abstract and provides little guidance on data collection. Thus, at the indicator level, the concept should be specific enough for empirical data to be gathered. Two important comments are in order. Both the evaluation and the research approach are adopted in the selection of indicators. The former requires information to be collected on intended, key aspects of the programme. The latter takes account of theoretically relevant explanatory factors. Taken together, the two sources constitute hypothesised causes. This choice of source has implications for the inclusion/exclusion of indicators at the third level. Furthermore, the relationship between indicators is substitutable since it reflects the reality of causal complexity better: *“both substitutability and equifinality stress that there are multiple paths to a given goal”* (Goertz 2006, 63).

Three indicators reveal the meaning of the dimension “condition substance”: the origin of the policy idea, political feasibility and technical feasibility. These three elements define ownership (Morrissey / Verschoor / 2004; Drazen / Isard 2004). As budget support conditionality emphasises the role of ownership in the effective achievement of development goals, it is important to verify whether the “new” conditionality integrates ownership elements and this approach produces better development results.

The first indicator asks about the origin of the condition substance, i.e. it belongs more on the donor side or the partner government side. The origin of the idea is important because experience of structural adjustment conditionality shows that unilaterally imposed policy changes are not implemented. The World Bank (World Bank 2006) and OECD (OECD 2006) guidance on budget support suggests that a policy change or target can be regarded as home-grown if it is derived from the PRSP. However, as Dijkstra (2011) demonstrates, a PRSP may be (i) dominated by donors’ agendas, (ii) disconnected from the real political process, being more of a means to obtain funds, or (iii) *“a weak link between the consultation processes and the actual writing of the strategy”* (116). An evaluator must therefore dig deeper into the negotiation process of the condition substance and the PRSP formulation process before he can identify its origin.

The next two indicators of the condition substance are political feasibility and technical feasibility. Political feasibility has two dimensions that together determine whether a policy

change takes place (Brinkerhoff 1996, 1396). First, political actors must engage in an intensive interactive process which involves “*consensus-building, participation of key stakeholders, conflict resolution, compromise, contingency planning, and adaptation*” (ibid). Second, policy implementation can alter the cost-benefit structure of both implementers and beneficiaries by reconfiguring roles and incentives. Drazen / Isard (2004) have highlighted the role of technical capacity as a key element in getting programmes executed. At times, technical capacity, e.g. the capacity to collect taxes, is crucial in implementing a programme. A demanding and laborious interactive process and potentially unfavourable cost-benefit distribution to key actors can be thus hypothesised as preventing reforms.

The finance component of the conditionality concept has two indicators that can effectively influence the likelihood of compliance with conditionality: predictability and financial stake. More predictable aid is an explicit aim of the Paris Declaration, and one which has been integrated into budget support. Predictable financial resources enable partner governments to plan and allocate resources across sectors in a more efficient and longer-term fashion. Predictability can thus be plausibly hypothesised as a key contributory factor in getting a reform programme executed. Two dimensions make up the predictability feature: a multi-year framework and a discrepancy between scheduled and disbursed aid (OECD 2011).

The second indicator in the finance component of the conditionality concept is the *financial stake*. In broad terms, the financial stake refers to the value that the partner government attaches to aid, relative to its access to other financial resources. In a principal-agent model, this factor may reduce the incentive to a government to implement its reform programme as the financial constraint eases (Killick 1997). Killick’s framework assumes a hard-core type of conditionality, which stands in stark contrast to the “consensual” type of conditionality that budget support donors promote. The inclusion of this factor enables evaluators to compare and verify the validity of different claims to effective models of conditionality, i.e. leverage versus consensus. This indicator may be joined by three other factors: aid dependency, dependence on budget support as an aid modality (ratio of GBS and balance-of-payment support to GDP–SPA survey series) and access to other sources of credit or financial support.

The last component of the conditionality concept, the link between condition substance and finance, is crucial to the definition of the conditionality mechanism at work. A credible sanction includes three factors: (i) the unambiguous nature of the condition’s content, (ii) a clearly perceived link between the fulfilment of a condition and the financial reward/sanction, and (iii) effective penalties.

The various aspects of the credible sanction allow evaluators to verify which (combination) of factors is more successful in obtaining the required result. Budget support donors differ in their approach to relaxing the link between funds and condition compliance. Experience from previous generations of conditionality shows that strong leverage conditionality cannot compensate for a lack of ownership in successful programme execution. In budget

support, therefore, innovation takes place not only in the condition substance component, but also in the linkage component.

An unambiguous condition is one that leaves little room for interpretation (OECD 2006, 32). The second aspect: a clearly perceived link between compliance and reward/sanction is crucial to the definition of the nature of conditionality. Here an evaluator can classify conditions as carrying substantial financial weight or not. BS donors differ in how they attach conditions to disbursement, the World Bank (WB) and a number of bilateral donors, for example, disbursing on the basis of a specific set of indicators. Another group of donors relax the link between funds and conditions to reflect their supportive and non-leverage approach to conditionality. In 2007 and 2008, for example, UK Department for International Development (DFID) did not attach any specific conditions to disbursement in Zambia, but based it on a general assessment of compliance with the underlying principles. It is not only useful to investigate which approach works better, but also necessary to verify one important, intentional aspect of a BS programme. The de-linkage feature of conditionality is well illustrated in the OECD guidelines on budget support. Donors should avoid bunching conditionality around a limited number of common criteria. The aim of this move is to prevent an excessive response to policy slippage and may therefore lead to the disruption of aid. Finally, the measurement of effective sanctions is straightforward: is non-compliance followed by a reduction of funds?

Several of the indicators have more than one dimension, political feasibility, for instance, consisting of interactive processes and cost-benefit distribution, credible sanctions and financial stakes both having three defining dimensions. One solution is to weight the various dimensions on the basis of the “*ontological theory of the object*” (Goertz 2006). In other words, weighting should follow what the theory claims to be more or less important. However, weighting is not possible if the variable is dichotomous. Another solution then is to keep the dimensions separate and measure them as separate indicators. Last but not least, a number of factors beyond the control of both BS donors and partner governments may influence the programme. For instance, such exogenous shocks as natural disasters and financial crises can lower the level of compliance with the conditions.

Matching evaluation questions with evaluation methods: a possible design for technical conditionality

The previous section outlined an analytical framework for BS conditionality at implementation level. The key questions at this level are how conditionality has been implemented (hypothesised causes), what the compliance level (effect) is and how the observed results are attributed to explanatory factors. This section will propose a possible research and evaluation design for these questions. It begins by briefly reviewing the available qualitative methods of establishing causality and goes on to justify the selection of Qualitative Comparative Analysis as a relevant method for these evaluation questions. The next step is to outline the application of the method.

The methods

Most qualitative methods for causal inference fall into one of two categories: cross-case analysis and within-case analysis. Cross-case analysis comprises comparative case study designs and Qualitative Comparative Analysis (QCA). QCA is based on Boolean algebra, and other tools in this family include multi-value QCA and fuzzy-set analysis. QCA does not “*manipulate numbers but rather systematizes logical expressions in order to sort data and create a list of the configurations of circumstances associated with a given outcome*” (Stokke 2007, 502). Its outstanding advantage is that it permits investigations of causal junctures or combinations of causes. The method possesses tools for the inclusion or discarding of explanatory variables in the search for factors substantially capable of explaining the outcome.

Comparative case study design has its root in Mill’s method of agreement and difference (Mill 1882). The method of agreement looks for similar antecedent conditions in two cases of similar outcomes. It is hoped that these will be necessary conditions. The most-different case study design follows the same logic as Mill’s method of agreement (Przeworski / Teune 1970). The investigator looks for cases in which independent variables differ (X2; X3; X4 assume different values) and co-variation exists only between X1 and Y1. This design hinges on a strong assumption that rarely holds in social science: the same cause holds across settings since outcomes are often the result of multiple causes (Gerring 2008, 671–674).

The most-similar case design reflects the logic of Mill’s method of difference:

“the investigator would look for antecedent conditions that differ between two cases that have different outcomes, and they would judge that those antecedent conditions that were the same despite differing outcomes could not be sufficient to cause either outcome” (Bennett 2004, 31).

In most-similar case design, an investigator would look for cases that display different outcomes, but whose explanatory factors have similar values. It is hoped that a thorough scrutiny of the cases will reveal one or more factors that differ between the two cases and can thus help to explain the outcomes.

Within-case analysis tools belong to the *realist* school of causality, which postulates causal relations according to actual events and processes that lead from causes to effects, rather than relying on regularity and correlations of events. The main thesis of the *realist* school of the realist approach to causality is that “*some causal processes can be directly observed, rather than only inferred from measured co-variation of the presumed causes and effects*” (Maxwell 2004). Causal process observation then becomes the main source of causal leverage for the qualitative method, and the cross-case section of the method plays a more “*supporting role*” (Goertz / Mahoney s. a.). Within-case analysis tools consist of pattern-matching, process-tracing and causal narrative (Mahoney 2000). Other tools for studying social processes include discourse analysis, discourse tracing and impact analysis.

Which method to use?

For evaluating the effectiveness of conditionality at implementation level, I propose a design which employs QCA as the main method for several reasons. First, QCA tools fit nicely with the evaluation task. The overriding question at this level is the question of attribution, i.e. to find out about the factors accountable for the observed level of compliance with conditionality (the compliance level is reported periodically by the partner government). The situation in which the evaluator finds herself is that, having measured a number of hypothetical causes and recorded the observed compliance level, the main question she has to answer is which causes or combination of causes determine the level of compliance with conditionality. QCA – based on the causal complexity concept – can identify the combination of causes responsible for the outcome. It detects and systemises patterns of causes and derives the causal configuration from the data. Second, the number of technical conditions per BS programme lends itself to QCA (30-40 for QCA, 40-50 for multi-value QCA and more for fuzzy-set analysis). Conducting intensive within-case analysis for all the conditions would be an almost impossible task, since the insights and wealth of information will be compromised if the number of cases increases significantly. Comparative case-study designs can handle only a few cases at a time and cannot capture conjunctions of causes for the entire population of cases.

A possible design

Case selection

The first step in conducting QCA is case selection. Two principles rule this step. First, cases should be comparable and demonstrate diversity of both hypothesised causes and outcome variables. Second, the contradiction/uniqueness problem should be borne in mind when the number of explanatory variables is included (Rihoux 2006). On the first point, an evaluator should make sure that the individual technical conditions are comparable in key respects. This is a challenge because, as pointed out in the previous sections, conditions vary according to policy fields and take different positions on the results ladder, from inputs to outcomes. As QCA uses dichotomous variables, similar values assigned to the same variable may mean different things in different cases as a result of this heterogeneity of the condition. For instance, positive political feasibility where the condition is the reform of state enterprises means a great deal of difference for positive political feasibility where the condition is the production of a report on the current state of women's land use rights. This issue leads to the infringement of the causal homogeneity assumption, which is a necessary condition if the causal inference is to be held valid: *“If two units have the same value of the key explanatory variable, the expected value of the dependent variable will be the same”* (King / Keohane / Verba 1994).

A solution to this problem is to “delimit” the population of cases to make them more comparable, using theory and empirical knowledge (Rihoux 2006). The challenge here is to choose the selection criteria so that an evaluator can constitute a population of cases that share similar background characteristics. Two such criteria might be the policy field and the level of policy outcome.

On the second point of selecting variables, if an evaluator includes too many explanatory variables, she will obtain a picture of full complexity with uniqueness of cases at the expense of generality and parsimony (Rihoux 2006). At the other extreme, when too few variables are included in the causal model, contradiction will increase (the same combination of factors produces different outcomes owing to missing variables). For a good balance a range of 4 to 6–7 explanatory variables for intermediate-N (10–40 cases) should not be exceeded (Rihoux / Ragin 2009).

An evaluator can use other techniques to select explanatory variables and specify the causal model better before conducting QCA. It should be noted that, as an approach, QCA permits confrontation and dialogue between the theory of the causal model and data. This point is well illustrated throughout different stages of QCA application. The most-similar and most-different designs can help to identify relevant explanatory variables. In the former design, an evaluator will look for cases that have as much similarity of explanatory variables as possible and yet produce different outcome levels (here level of compliance with conditionality). Once the cases are “controlled”, the evaluator will look for one or more factors that assume different values in these cases, enabling her to attribute the difference outcomes to them. She can then include these factors in the QCA analysis. For instance, evaluators may detect two or more cases of reform that are highly feasible technically and receive solid support from influential domestic groups and strong financial backing from donors. Despite these shared features, some aspects of the reform are successful, while others fail. Evaluators will then look for one or more factors that account for the performance and include them in the evaluation framework. Similarly, the most-different design allows an evaluator to select relevant explanatory variables and eliminate irrelevant ones by examining cases that share similar outcomes but maximum heterogeneity of explanatory factors. The evaluator will then look for and retain one or more similar factors that may account for the outcome and include them in the QCA analysis (Rihoux 2006).

Measurement and dichotomisation

In this step, the evaluator studies her cases and assigns values to explanatory and outcome variables. As QCA works with dichotomous variables (0;1), she should take extra care in coding since even modest modifications to operationalisation can give rise to profound changes in the final result (ibid). The greatest risk posed by dichotomisation is data loss. Rihoux / Ragin (2009) develop several guiding principles for this step. First, transparency is needed in justifying the threshold (when to assign a case to the [0] or [1] value), and it must be justified on the basis of empirical and/or theoretical knowledge. For example, if the data show that most cases display either of the tendencies: funds are disbursed according to selected, known individual conditions, they are disbursed according to the performance level of all conditions, or they are linked not to compliance with technical conditions, but to the underlying principles. The evaluator can then assign cases of the first two disbursement mechanisms to [1] (meaning there is a clear linkage between condition compliance and fund) and cases of the third mechanism to [0] (meaning there is no clear linkage). The second solution is to transform multi-category nominal variables into binary variables. If

the evaluator already possesses fine-grained data (interval or ordinal), she can turn them into multi-category nominal variables and then transform them into binary variables.

The last two important instructions given by Rihoux and Ragin concerning measurement and dichotomisation are (i) “*avoid artificial cuts dividing cases with very similar values*” and (ii) “*code the conditions in the correct ‘direction’, i.e. those coded [1] (positive, presence of the hypothesized causes) should be theoretically related to the positive outcome*”. However, dichotomisation should not be conducted at all costs, as some phenomena are non-dichotomous by nature. Forcing such factors into dichotomous variables will create “*problematic measurement bias*” (Braumoeller / Goertz 2000, quoted by Rihoux 2006). The solution here is to move to multi-value QCA or fuzzy-set analysis. Fuzzy-set analysis allows for membership in the interval from 0 to 1 depending on the actual state of the case and thus acknowledges the dimensions of cases as they actually are (Ragin 2000).

Data analysis: truth table

After studying each case and assigning values to explanatory tables and outcomes, the evaluator will be able to compile a raw table that shows that each case has a specific combination of explanatory factors (of value 0 or 1) and outcomes (of value 0 or 1). A software programme will then produce a truth table that displays a list of configurations showing how combinations of factors lead to outcomes. The next step is to minimise the long expression in the truth table to the shortest possible expression to detect regularity. Lastly, an interpretation is made on the final minimal formula in terms of the causal model (Rihoux 2006, 683).

Table 1 is a hypothetical truth table relating to the level of compliance with technical budget support conditionality. There are 23 cases and 4 hypothesised causes or conditions. The truth table reveals four routes to success and two to failure:

$$\text{Compliance} = AbCD + aBCd + abCD + abCd \quad (1)$$

$$\text{Non-compliance} = AbcD + ABcD \quad (2)$$

These two equations are descriptive in that they do not go beyond expressing the state of the empirical data. The next step is to minimise these long expressions by removing redundant conditions and producing shortest possible expressions that capture the regularity of data (Rihoux 2006). The rule for Boolean reduction is as follows:

“If two Boolean expressions differ in only one causal condition yet produce the same outcome, then the causal condition that distinguishes the two expressions can be considered irrelevant and can be removed to create a simpler, combined expression”
(Ragin 1987, quoted by Rihoux 2006).

In equation (1), both $AbCD$ and $abCD$ lead to compliance. Uppercase letters denote a [1] value (“*large, presence, high...*”), while the lowercase letters refer to the [0] value (“*small, absence, low...*”) of the binary variable (Rihoux / Ragin 2009). We see that, no matter

which value condition A assumes (0 or 1), the positive outcome still occurs. These two expressions can therefore be minimised to bCD . By the same reasoning, the remaining two paths, $aBCd$ and $abCd$, can be reduced to aCd , since compliance occurs no matter which value variable B assumes.

Thus equation (1) can be reduced to:

$$\text{Compliance} = bCD + aCd$$

and finally: $\text{Compliance} = bC + aC$ (3)

The path to failure, or non-compliance, is expressed in equation (2) and can be minimised as AcD , since condition B is superfluous in causing conditionality compliance.

Conditions				Outcomes	Number of Cases
A	B	C	D		
1	0	0	1	0	2
1	0	1	1	1	5
0	1	1	0	1	7
0	0	1	1	1	4
1	1	0	1	0	2
0	0	1	0	1	3

A = origin of policy idea; B = political feasibility; C = financial stake; D = credible sanction; outcome = condition compliance
Source: own compilation

Once the final configuration or the *prime implicant* is produced, the evaluator can interpret. At this stage, there is an opportunity to compare empirical reality with the theoretically derived hypothesis. For example, the prime implicant for success (equation 3) shows that C, the financial stake, is necessary for compliance with conditionality because it is always present when the outcome occurs. In this hypothetical case, the financial stake can – contrary to common belief – compensate for a lack of political feasibility and origin of the policy idea. The prime implicant for failure (AcD) confirms this finding: compliance with conditionality is not achieved in the absence of the financial stake, and the presence of a credible sanction or ownership of the policy idea cannot compensate for its absence.

One important note on interpretation: evaluators should not single out individual conditions in relation to outcomes, but interpret them as a configuration or intersection of factors. Embracing this principle will help evaluators to take advantage of the strength of QCA: detecting causal complexity (Rihoux / Ragin 2009). Evaluators should single out conditions

only when they are (or come close to) necessary and sufficient conditions, as in this hypothetical example.

This section has provided a simplified version of steps in conducting QCA. In practice, evaluators are most likely to be confronted with contradictions in the process. Contradictory configuration signals that there is something to be learnt from the cases and should be considered part of the “*interactive process of ‘dialogue between ideas and evidence’*” and not as a failure (Ragin 1987, quoted by Rihoux / Ragin 2009). The next section will deliberate on the process of dealing with contradictory configurations.

Dealing with contradiction

Contradictory configurations are those that are “*counterintuitive*”, for example configurations in which all [0] variables lead to [1] outcomes or vice versa (ibid). Another type of contradiction arises when the same configuration leads to contrasting outcomes. Rihoux and Ragin list eight solutions to resolve contradiction. Only selected solutions are discussed here; for a full version see Rihoux / Ragin (2009, 48–49).

- Conceptualisation and operationalization of explanatory factors: “*reexamine the way in which the various conditions included in the model are operationalized. For instance, it may be that the threshold of dichotomization for a given condition is the source of the contradiction between two cases*” (ibid). For example, an evaluator may go back to contradictory cases and find that political feasibility in some cases means the support of reform implementers, but in other cases the support of top politicians. The necessary task then is to reconceptualise political feasibility.
- Conceptualisation and operationalization of outcomes: “*reconsider the outcome variable itself ... If the outcome has been defined too broadly, it is quite logical that contradictions may occur*” (ibid). For instance, compliance with conditionality can mean non-compliance, compliance and partial compliance. An evaluator may reconsider how thresholds that determine whether partial compliance falls under non-compliance or compliance. Alternatively, a technique for dealing with this problem is to turn multi-category nominal variables into several binary variables.
- Missing variable: “*re-examine, in a more qualitative and ‘thick’ way, the cases involved in each specific contradictory configuration. What has been missed? What could differentiate those case, that hasn’t been considered, either in the model or in the way the conditions or outcomes have been operationalized?*” (ibid). An evaluator may find that, for certain configurations of explanatory factors, conditionality may or may not be complied with (the configuration appears in both compliance and non-compliance equations). It is to be hoped that, by going back and studying the cases in detail, the evaluator will find one or more factors that can explain and thus differentiate the cases. For example, two cases of reforms may receive high levels of support from key domestic actors, have low technical feasibility and face credible sanctions for non-compliance. Yet one case is a clear success, while the other is an utter failure. The evaluator may then look for any factors responsible for the difference in performance that have not been taken into consideration before.

Concluding remarks

This section has demonstrated how QCA can be used to evaluate the effectiveness of BS conditionality at implementation level. It should be remembered that the evaluation questions at this level are: (i) how has conditionality been delivered, (ii) what is the compliance level and (iii) what factors account for the success or failure of compliance? In a nutshell, QCA is capable of answering the evaluation questions, with the following advantages and challenges:

Advantages:

- The *method helps to answer the evaluation question*. The measurement of explanatory variables captures the main features of how conditionality is implemented. Moreover, it tells us which aspects of implementation matter for the compliance level. An evaluator is free to include variables that constitute the core theory of the BS programme and should be verified, and those that she deems relevant in explaining the compliance level. At the end of the process, the evaluator will have obtained prime implicants that indicate conjunctions of factors accounting for compliance or non-compliance with conditionality.
- *Generalisation is gained, while proximity to cases is not lost*. With the qualitative Comparative Analysis method, an evaluator can work with a medium number of cases. She can thus grasp the entire picture of technical conditionality. This is an advantage, since other “traditional” qualitative methods allow analysis of only a handful of cases at a time. More importantly, this generalisation is gained without compromising the knowledge of particular cases. An evaluator needs first to gain a reasonable understanding of cases before she can code the data and conduct analysis. The conjunctions of factors that account for compliance or non-compliance with conditionality are therefore extracted directly from the cases. This key strength has been summarised as follows: “*by using QCA, the researcher is urged not to specify a single cause model that fits the data best, as one usually does with statistical techniques, but instead to determine the number and character of the different causal models that exist among comparable cases*” (Ragin 1987, quoted by Rihoux 2006, 681).

Challenges:

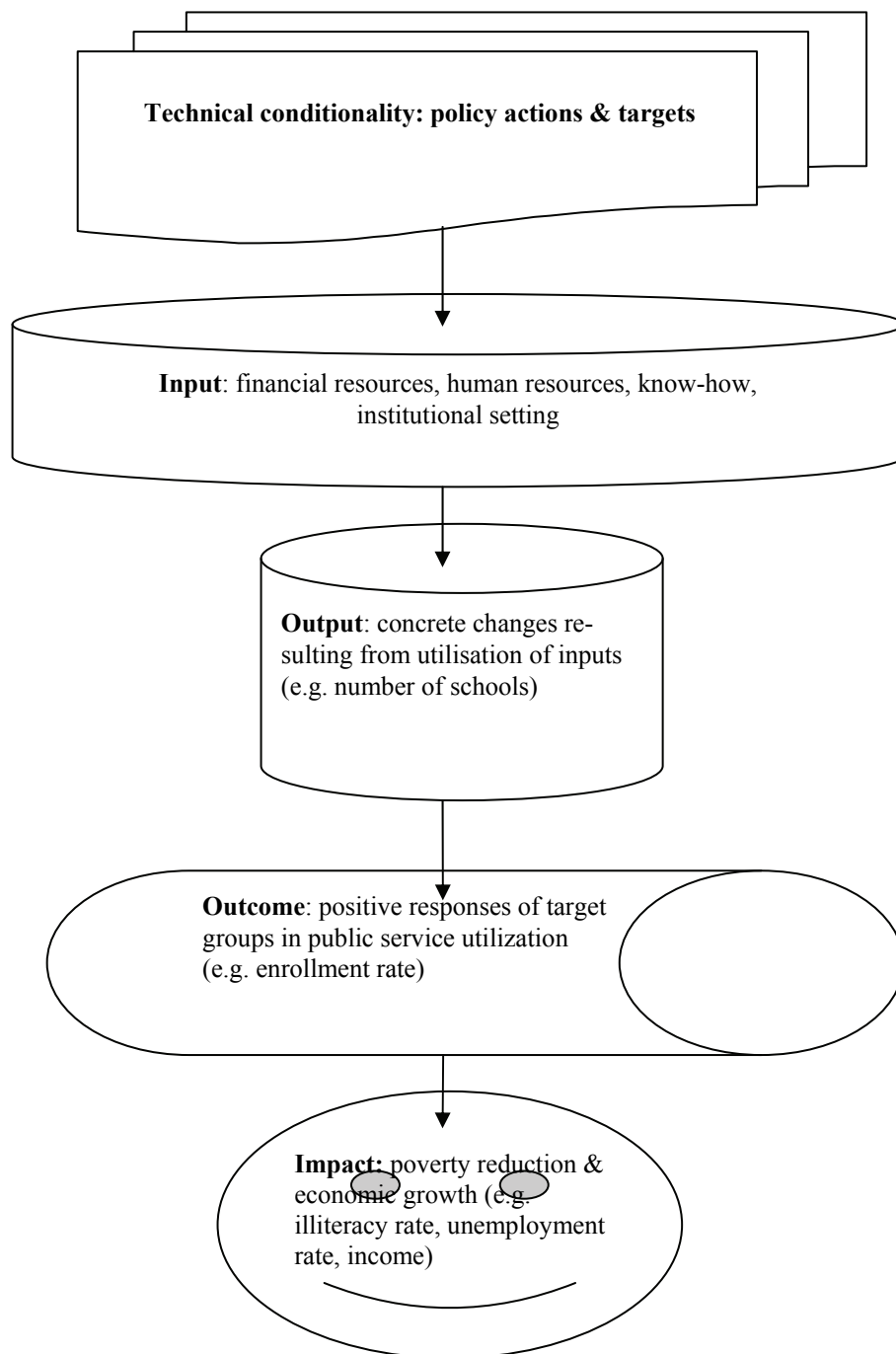
- *Dichotomisation as a potential source of data loss*. This is the most commonly cited critique of the method. An evaluator should consider what distinguishes the cases she studies, whether the difference is of *nature* or *kind*, or difference of degree (Ragin 2002, quoted by Rihoux / Ragin 2009). At this stage of the discussion, dichotomisation is a challenge, but not a problem for the evaluation of BS conditionality. An evaluator can determine fairly accurately in each case whether the financial stake is high or low, the policy idea is the partner government’s rather than the BS donor’s, technical capacity is strong or weak, etc.
- *Dealing with causal heterogeneity*: a precondition for comparing cases is that the cases are comparable. In other words, a comparison of “apples and oranges” should be avoided. Putting non-comparable cases together will create the problem of two units of

the same value of the explanatory variable resulting in the dependent variable having a different value. The challenge here is to select criteria so as to establish a sub-set population of cases which share sufficient features to make comparison meaningful. Another solution is expressly to acknowledge the difference that is causally related to the outcome, to measure it and to include it in the analysis.

Evaluating the effectiveness of conditionality: an analytical framework

This section considers the main aspect of the evaluation of the effectiveness of BS conditionality. Its structure is similar to that of the previous section: it begins by envisaging an analytical framework and then proposes a research and evaluation design. The analytical framework raises the question of what to measure, or what are the hypothesised causes and effects. The nature of the intervention itself (technical conditionality) and the theory of programme evaluation justify the selection of hypothesised causes and effects.

First, the nature of the intervention is such that technical conditionality consists of a number of highly diversified individual conditions, each with its own intervention logic. These various conditions can be differentiated by reference to two dimensions. The first dimension is the policy field. If we categorised technical conditions according to policy fields, we would end up with many types, from agriculture to transportation, health, education, environment, public financial management, etc. The second dimension is the level of changes. Figure 2 summarises the key steps in the “production” of public goods and services, i.e. input, output, outcome and impact. Technical conditions require changes to take place at all levels, except impact. Implementation evaluation enables evaluators to measure and explain compliance with technical conditions regardless of how they differ from one policy field and level of changes to another. However, there is no such common denominator, since each condition has objectives that differ in content and level of changes.

Figure 2: Technical conditionality

Source: own illustration

Box 1: Example of technical conditions

Input: revise tuition fees at secondary and tertiary levels, to reflect market conditions better, and enhance policies to protect the poor (Viet Nam PRSC 7)

Output: Number of hectares of land newly brought under irrigation and of land with old irrigation systems rehabilitated (Agriculture sector, Zambia PAF 2008-2010)

Outcome: percentage of fully immunised children under one year of age in 20 worst performing districts (Health sector, Zambia PAF 2008-2010).

Second, the theory of programme evaluation suggests that evaluating the *effectiveness* of a programme means selecting the cut-off point at outcome level. MacLaughlin and Jordan have defined evaluation as the “*systematic investigation of the merit or worth of an object for the purpose of reducing certainty in decision making about that object*” (McLaughlin / Jordan 2004, 27). To evaluate the merit of a programme is to ask whether the programme has achieved its aims, while to evaluate the worth of a programme is to focus on the broader impact. One addresses the issue of merit by examining whether a programme has achieved its short- and medium-term outcomes. A programme’s influences and effects are strongest and most visible at this level. To evaluate the impact, evaluators must study the longer-term outcomes, which interact with other forces to produce changes. Consequently, evaluating the effectiveness of a programme means investigating whether changes occur at outcome level, where “*the program sphere of influence usually stops*” (ibid).

Box 2: OECD DAC definition of the evaluation of effectiveness

“Effectiveness

A measure of the extent to which an aid activity attains its objectives. In the evaluation of the effectiveness of a programme or a project, it is useful to consider the following questions:

- To what extent were the objectives achieved/are the objectives likely to be achieved?
- What were the major factors influencing the achievement or non-achievement of the objectives?”

(DAC Criteria for Evaluating Development Assistance, OECD 2000)

Two important points emerge when we draw together the two sources to construct an analytical framework for evaluating the effectiveness of BS conditionality. First, beyond the implementation level, each condition has its idiosyncratic intervention logic with hypothesised causes and effects. This stems from the fact that conditions differ in both content (policy field) and the level of change that is their aim. This leads to the second point: the analytical framework for individual conditions must be constructed because of their distinct intervention logics. Most of the time, the intervention logic can be found in the budget support donors’ programme documents. Box 3 gives examples of a goal matrix of conditions in the fields of education, public sector management and governance. However,

when a BS programme document says nothing about the aims/objectives of technical conditions, evaluators must begin by reconstructing the goal hierarchy of those conditions.

To make the evaluation of an intervention possible, an analytical framework far beyond what we find in Box 3 is needed. In other words, a goal hierarchy is necessary, but not sufficient for the task of evaluation. While BS donors could provide evaluators with brief outlines of their intervention logic, evaluators would have to devise an analytical framework based on this preliminary information. The higher aims and objectives of the individual conditions are often kept at an abstract level. An evaluator should develop a profile of the aims/objectives consisting of the key dimensions so that she can assess whether change has occurred. Omitting this step will fundamentally weaken the causal inference and attribution at a later stage. For example, evaluators of condition (2) in Box 3 would have to identify the features of the prosecution process that must be there for her to conclude that cases arising out of a Public Accounts Committee report had been satisfactorily prosecuted. An evaluator also needs to make clear which aspects of the legal and institutional framework have been strengthened as a result of these cases being successfully prosecuted. Finally, at policy objective level, she should indicate which types or relationships of public accountability have been improved.

Governance effects

Two compelling reasons justify the need to evaluate the governance effect of BS technical conditions *at the same time* as its effectiveness. First, improving governance issues is one of the major objectives of BS donors. Germany, for example, has governance as one of its three objectives in providing BS, the others being effectiveness/efficiency and financing (BMZ 2008). At the core of BS intervention logic stands the notion that aid effectiveness cannot do without the enhancement, rather than the bypassing, of the governance system. The transfer of financial resources directly through the national financial system and concurrent efforts to build up the institutional capacity of this system exemplify the governance agenda embedded in budget support.

Box 3: Examples of the goal hierarchy of technical conditions			
Condition	Expected outcomes	Policy objective	Source
1. Improve the quality of learning through in-service training on the new curriculum (30% of teachers) and supply of textbooks for core subjects to all primary schools	Teacher performance improved alongside pupil learning achievement and outcomes (literacy and numeracy)	Improve quality of teaching and learning	Ghana MDBS Policy Matrix 2008-2010
2. Begin prosecution of cases arising out of 2004/5 Public Accounts Committee (PAC) report	Legal and institutional framework to reduce fraud and efforts to combat corruption strengthened	Improve governance and public accountability	Ghana MDBS Policy Matrix 2008–2010
3. Submit to Cabinet a proposal for a comprehensive public-sector pay reform (including the new pay range structure and budgetary implications) 30% (PRSC08) and 50% (PRSC 09) of public servants remunerated on the basis of a unified national pay spine	Improved equity in public-sector pay structure through establishment and use of a new pay structure to guide remuneration	Increase the capacity of the public and civil service for accountable, transparent, timely, efficient and effective performance and service delivery: pay reform	Ghana MDBS Policy Matrix 2008–2010

Second, technical conditions, in the form of policy actions and policy targets, hardly escape governance effects. These policy changes do not take place in an institutional or political vacuum. On the contrary, they are the “new arrivals” in a well-established institutional landscape. Institutions – formal (hierarchy of laws and regulations) and informal (norms, conventions, codes of conduct, traditions) – govern human exchanges and decide what is possible (North 1991). Human agency is both affected by the institutional constraints and actively seeks to alter the rules of the game as well as its outcome to its own advantage, through a patronage network, for example. A development intervention, be it a programme, a project, a policy measure or a target, is inevitably governed by existing rules and games. It is thus appropriate to investigate the interaction between BS conditionality and its immediate institutional and political economy environment. However, causality is not a one-way street. Local institutions and human agencies are as likely to influence the feasibility and final shape of the proposed policy changes as that intervention is to seek to “improve” local institutions and the way they and human agency interact. Evaluators should therefore check complex interaction processes for mutual causality and unintended effects.

Finally, we come to the practical question of defining an analytical framework for the effects of BS technical conditionality on governance. The construction of this framework should comply with two principles. First, evaluators should identify hypothesised

Box 4: Example of governance effect profile for a technical condition

Intervention (technical condition): prosecute cases arising from the State Audit Report

One class of case: public investment projects are initiated not on economic or social grounds. Rural markets are used one or two months each year, irrigation schemes cannot carry water to the fields because such important items as secondary channels are missing, etc.

Crucial governance elements: weak legal accountability: public managers not taken to court for misconduct; extensive cronyism and patronage: lucrative construction contracts awarded to well-connected companies and generous kick-backs paid to public project owners determine investment decisions; weak political accountability: members of parliament do not represent people's interests or question the misconduct of public managers.

Unit of analysis: individual, network

Hypothesised governance effect (if the cases are prosecuted)

At implementation level: patronage networks are alarmed and possibly weakened, a decrease in such malpractice as dividing bid contracts into smaller package to avoid competitive bidding

At outcome level: public trust and confidence boosted and possibly lead to greater voluntary compliance with the law

At policy objective level: legal accountability is strengthened

governance effects for *all* levels of the goal hierarchy. The central idea is to devise a tentative map of governance issues around any policy action/target and its strategic objectives. Second, evaluators should select the unit of analysis in accordance with the level of change that is the aim of the policy action/target.

Evaluating the effectiveness and governance effects of technical BS conditionality: a possible design

Which method to use?

Of the qualitative methods available for the investigation of causality, within-case analysis and cross-case analysis, the former is better at evaluating the effectiveness of BS technical conditionality and gauging its effects on governance, mainly because of the number of cases and observations of technical BS conditions. Cross-case analysis involves the investigation and comparison of a sample of cases, while within-case analysis entails the intensive study of individual cases. A case is defined as a "*unit of analysis ... about which information is collected and inferences are made*" (Brady / Collier 2004, 229). A case may contain one or more observations, which have several dimensions, each dimension often being measured as a variable. For example, when evaluating the effectiveness of BS conditionality, we are interested to see if each and every condition achieves its intended objectives. As such, each condition is a *case*, and we would like to know, for example, whether the implementation of the recommendations made in the State Audit Report results in the improvement of legal accountability. The previous section argued that each BS technical condition has its own intervention logic, with idiosyncratic causes and effects. Each condition therefore constitutes one case. In other words, there are no other cases sufficiently similar in important respects (causal factors and intended effects) for

comparison. Consequently, each BS technical condition lends itself to within-case analysis when the aim is to investigate its effectiveness and governance effects.

One outstanding feature of within-case analysis is that it makes causal inferences from causal process observation, not data-set observations. It has several tools, but this paper chooses to work with process-tracing, owing to its proximity to the task of causal process observation. In process-tracing, the steps that connect causal factors with outcomes are identified and the implications of the causal mechanisms are verified. It takes account of the role of contexts and temporal and spatial orders in building a credible causal link. Importantly, causal mechanism should not be equated with intervening variables, which takes the approach back to the correlational assumptions (Mahoney 2001, quoted in Falleti / Lynch 2009):

“...mechanisms are relational concepts. They rise above and outside the units in question, and they explain the link between inputs and outputs. Mechanisms describe the relationships or the actions among the units of analysis or in the cases of study. Mechanisms tell us how things happen: how actors relate, how individuals come to believe what they do or what they draw from past experiences, how policies and institutions endure or change ...” (Falleti / Lynch 2009, 1147)

Thus the application of process-tracing naturally requires an understanding of the causal mechanism at work. The next sections will discuss case selection and data analysis and draw conclusions on the capacity of the method to answer the evaluation questions.

Case selection

Qualitative and quantitative methods are used to select cases for study on different grounds and for different purposes. As cases are selected randomly in quantitative study, the result of the study can be confidently generalised to the population of cases from which the sample is drawn. In qualitative methods, cases are selected because they are information-rich and capable of producing an in-depth understanding. The purpose here is to gain insights and to shed light on the main question being studied (Patton 2002, 230). It is perhaps useful to reiterate the evaluation questions when cases for investigation are being chosen:

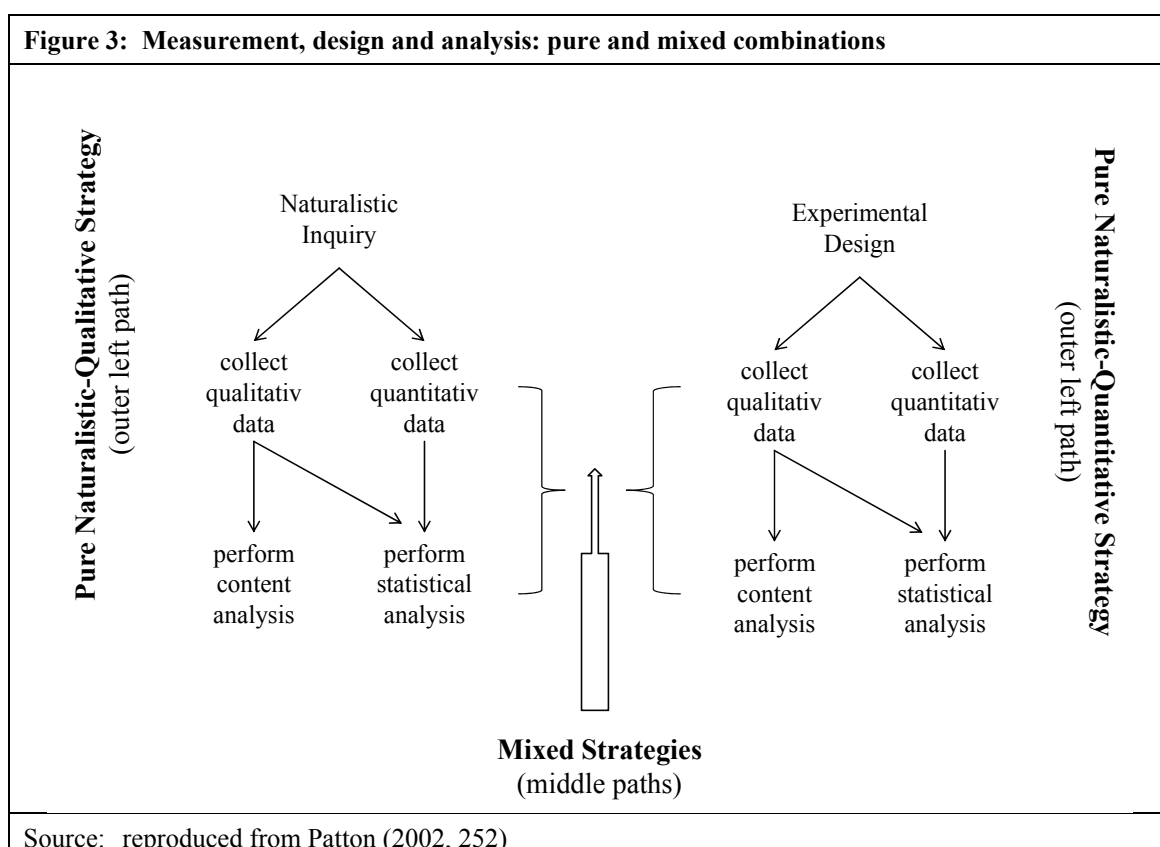
- To what extent were the objectives achieved/are the objectives likely to be achieved?
- What were the major factors influencing the achievement or non-achievement of the objectives?
- What are the governance effects?

As this set of questions applies to each and every BS technical condition, and each condition is a unique case, evaluators should in principle examine all cases. However, two strategies may help to reduce the number of cases to be studied. First, if evaluators wish to know the effectiveness of conditionality, she should focus on positive cases only, i.e. cases in which the conditions are complied with. The effectiveness of a certain intervention simply cannot be evaluated if that intervention is not or only partially implemented. In such

cases, evaluators can go back to the implementation level and inquire into the reasons for non-compliance (see the previous section).

Once the evaluator has eliminated negative, non-compliance cases, she is still left with numerous cases and may consider the second strategy for further reducing the number to a reasonable level. Evaluators can group cases according to their causal mechanisms and select from each group one or two typical cases for in-depth study. Although the individual conditions are moderate in number (30–100), probably less than a handful of mechanisms are at work. However, evaluators should have a good technical knowledge of types of causal mechanisms and a reasonable amount of information on the causal mechanism at work in each case. Causal mechanisms may exist at macro or micro level, and the causal agents may be individual, collective or social systems (Falleti / Lynch 2010). The World Bank, for example, very often has as its prior actions enactment of new legislation in various policy fields, such as regulations on the disclosure of state audit reports, press laws on the reporting of corruption and laws on the central bank's mandate and autonomy. These seemingly highly distinctive laws share a common attempt to prompt change through positive feedback, i.e. to create new constituencies with incentives to change the policy agenda.

Design, data collection and data analysis



Evaluators can select a purely qualitative, a purely quantitative or a mixed design, as shown in Figure 2. The degree of control in the inquiry (naturalistic or experimental), the type of data and the method of data analysis form the backbone of a design strategy. For a number

of reasons, a mixed strategy is the best choice for evaluating the effectiveness of BS technical conditionality. First, programme evaluation entails the systematic collection of information on programme activities, certain characteristics and different levels of outcomes. Evaluators consequently need to adopt a hypothetical-deductive approach, which is often associated with experimental, quantitative methods. The hypothetical-deductive approach requires the specification of variables and hypotheses *before* data collection. At the data analysis stage, information and data will be analysed and arranged in accordance with the existing framework.

The first step in designing an evaluation of the effectiveness of BS technical conditionality is therefore to construct a framework that guides data collection. The previous section provided guidance on how to select the intervention and the effects of technical conditionality and its governance effects. However, for process-tracing, a framework that connects the path from intervention to the hypothetical effects is needed. For example, the intervention logic may identify the improvement of legal accountability as an effect of the prosecution of cases arising from the Public Audit Report. The evaluator should then build frameworks that describe the main actors involved, the incentive structure that governs their relationships and how the prosecution of this particular case weakens or alters the nature of those relationships. Finally, she should make clear how the change in the relationships among key actors may improve legal accountability, public trust and voluntary compliance with the law.

The construction of an analytical framework for process-tracing should comprise three key steps. As **the first step**, evaluators examine the *context* in which the intervention functions. The context can be at macro, meso or micro level. What are important are the elements of contexts that have a direct bearing on the operation of the intervention. The context is important because it mediates the influence of causal mechanisms:

“credible causal social scientific explanation can occur if and only if researchers are attentive to the interaction between causal mechanisms and the context in which they operate unless causal mechanisms are appropriately contextualized, we run the risk of making faulty causal inferences” (Falleti / Lynch 2009, 1144).

For example, the WB’s conditionality on legislative reforms in many policy fields typically aims at fostering a new type of policy constituency by legitimising its voice and role or altering its mandate. Evaluators should therefore examine how weak legal accountability and strong patronage networks, which are ubiquitous in developing countries, may influence these proposed policy changes.

As the second step, evaluators should identify key actors, their motivations and perceptions and the micro-foundation causal mechanism at work. To continue with the example of law enactment as a common type of World Bank conditionality, one of the prior actions of PRSC 8–10 for Viet Nam reads: *“Prepare a revised Press Law to encourage accurate, objective and responsible reporting on corruption”*. The key actors are the press community, government officials and the public. The micro-foundation causal mechanism here is policy feedback, in which “new policies create new politics”.

Legal clarity and thus the protection of reporting on corruption can be expected to motivate the press to expose more cases of corruption. Public pressure and the prosecution of such cases lead to the adoption of rules designed to prevent corrupt practices. Politicians and bureaucrats are thus constrained institutionally and discouraged from committing corruption. Public accountability increases as a result.

As the third step, evaluators should formulate hypotheses or predictions on the basis of the hypothesised mechanisms. This step is necessary for the later stage of data analysis, when evaluators conduct and interpret test results (Collier 2010). Those predictions should focus on critical elements that lead to success or failure and, at the same time, cast light on the dynamic of relationships among key actors and the operation of those relationships in practice.

As an example of the third step, evaluators might propose a series of changes that must be observed if the revised press law is to result in an increase in public accountability through the causal mechanism of positive policy feedback. The predictions might be as follows:

- An increase in the number and scale of corruption cases exposed by the press *after* the promulgation of the law
- The majority of cases and/or “big” cases are prosecuted
- Policy learning takes place in various forms:
 - An increase in bidding practices in which transparent procedures are adopted
 - A decrease in bidding practices in which competitive bidding is avoided through the manipulation of the law

Data analysis

Descriptive inference “*Fine-grained description is a foundation of process tracing*” (Collier 2010, 4). Indeed, description is the “bedrock” of qualitative reporting, be it in research or programme evaluation (Patton 2002). The key issues in description are the analysis and organisation of data. Three main approaches in qualitative reporting are chronology, case study and analytical framework (ibid). In the chronological approach, an evaluator can organise information in chronological sequences, which may extend from the beginning until the end (history) or from the outcome backwards to the beginning (flashback). Evaluators may also choose to report their data and information in the form of a case study. The central unit of analysis may be people (individuals or groups), critical incidents that shape the outcomes or settings (places, sites, locations). In programme evaluation, evaluators can build layered or nested case studies, which consolidate smaller units in a larger one when a programme is implemented at different administrative levels (Patton 2002, 439).

Finally, evaluators can organise data and information in accordance with an analytical framework. Here data are described along important processes, e.g. programme formulation, key decision-making, implementation, etc. Alternatively, evaluators can

arrange data on the basis of relevant issues or report in response to questions, especially if a standardised format has been adopted for the interviews (ibid).

Evaluators can adopt an *inductive* or *deductive* approach to analysis to organise raw data into meaningful description. In both approaches, they will conduct “*pattern, theme and content analysis*” to categorise information. This step involves “*identifying, coding, categorizing, classifying and labelling the primary pattern in the data*” (Patton 2002, 463). In the inductive approach, evaluators let the themes and patterns emerge from the data, while, in the deductive approach, they report in accordance with a framework.

In this particular case of using process-tracing to evaluate the effectiveness and governance effects of BS conditionality, I propose that the analytical framework approach should be employed. However, evaluators should take the middle road and be aware of new developments and findings outside the framework and incorporate the new issues in their reporting. They should treat the framework as a useful starting point for data collection when tracing the causal process. However, they will probably encounter deviations, uncover other mechanisms at work, detect unintended effects resulting from interaction with contexts, etc. Evaluators thus start from the analytical framework, but should be open to the development of the case in reality, follow the trail and accurately record the real-life effects and mechanisms.

A good example of descriptive inference can be found in Geske Dijkstra’s work on Poverty Reduction Strategy Papers (2011). Dijkstra describes the PRSP formulation process in three countries: Bolivia, Nicaragua and Honduras. The author identifies the key actors, often ministries, committees, parliaments, civil society organisations, WB, IMF, multilateral and bilateral donors. The author then follows sequences of events, pinpoints the motivations of key actors and analyses the content of PRSP documents. By putting together different pieces of evidence at different levels of analysis, Dijkstra reconstructs the process of how actors relate to one another when drawing up a PRSP and how this process has direct implications for the result of a PRSP.

Dijkstra’s work achieves the objective of process evaluation, which is to extract from the “details and dynamic” of the programme process individual factors to which the success or failure of a programme can be attributed (Patton 2002, 160). In Bolivia, for example, the author highlights the disconnection between the consultation process and the actual drafting of the PRSP. At the drafting stage, participation is restricted to sector ministries and the donors, and it is possible to attribute different aspects of the PRSP to the influence of either side. The weak link between consultation and drafting, combined with evidence of the donors’ influence on the PRSP content, helps the author to draw the credible conclusion that the PRSP does not reflect national ownership:

“When interviewed in April 2003, many donor representatives considered that the EBRP suffered from a number of limitations, such as a lack of prioritisation, with a heavy focus on the social sectors and a lack of attention to growth, it was too technocratic, with no attention being paid to political feasibility, and there was a lack of domestic ownership. However, given the process as described above, these weaknesses can to a large extent be ascribed to excessive donor influence. Donors wanted to include all their existing

projects, they focused heavily on the social sectors, and they preferred to work at a technical level and tended to disregard politics. The result was clearly a lack of ownership” (Dijkstra 2011, 116).

Evaluators should not least try to achieve “internal homogeneity” and “external heterogeneity” while classifying information. The former principle requires that information and data within the category be connected in a meaningful way, while the latter indicates substantial differences among categories.

Causal inferences: the four tests

Qualitative data analysis is the stage at which an evaluator/researcher “*transforms data into findings*” (Patton 2002, 432). Consequently, data analysis very often ends at the description stage, and insights into causality may result from this process. Qualitative researchers and evaluators perform outstanding qualitative work and draw sound causal inferences, despite the absence of hard and fast rules (e.g. Theda Skocpol’s study on social revolution, Henry Brady’s evaluation of the Florida election results in the 2000 presidential election, Dijkstra’s evaluation of the PRSP formulation process, etc.). As within-case analysis is the main source of causal leverage in small-N qualitative study, scholars have recently developed more stringent procedures for causal inference. The remainder of this section will introduce the test and give some examples. While this test makes the basis and logic for drawing causal inferences more explicit, the backbone of these tests and thus of qualitative data analysis continues to be description.

Hoop test. This test provides a “*necessary but not sufficient*” criterion for discarding or accepting a hypothesis. In other words, the hypothesis must pass this test to remain valid. However, the test is not sufficient for the acceptance of a hypothesis. Passing this test alone does not therefore lead to affirmation of the hypothesis (Bennett 2010). To revert to the example of the effect of the prosecution of cases referred to in the State Audit Report on legal accountability, it must be possible to observe that a certain number of cases have been prosecuted. If none of the recommended cases are prosecuted, evaluators can safely conclude that there has been no effect on legal accountability. However, the prosecution of cases alone does not guarantee that legal accountability has been improved.

Box 5: Process tracing: four tests for causation			
Sufficient to establish causation			
Necessary to establish causation	No		Yes
	No	<p>Straw in the wind</p> <p><i>Pass:</i> affirms relevance of hypothesis, but does not confirm it</p> <p><i>Fail:</i> suggests hypothesis may not be relevant, but does not eliminate it</p>	<p>Smoking gun</p> <p><i>Pass:</i> confirms hypothesis</p> <p><i>Fail:</i> does not eliminate it</p>
	Yes	<p>Hoop</p> <p><i>Pass:</i> affirms relevance of hypothesis, but does not confirm it</p> <p><i>Fail:</i> eliminates hypothesis</p>	<p>Doubly decisive</p> <p><i>Pass:</i> confirms hypothesis and eliminates others</p> <p><i>Fail:</i> eliminates hypothesis</p>
	Source: Bennett (2010, 211)		

Smoking gun test. This test provides a “sufficient but not necessary” criterion for discarding or accepting a hypothesis (ibid). If evaluators can present evidence in support of a certain hypothesis, they can affirm it. However, the absence of that evidence does not help them to discard the hypothesis. For example, if it is seen that the practice of intransparent bidding decreases, it can be concluded that legal accountability has been strengthened. However, the absence of that evidence does not mean that there has been no change in legal accountability, since there is more than one way to improve it. For instance, if courts take action to prosecute cases in a fair, effective and efficient way, and public trust increases, it can still be safely concluded that legal accountability has improved.

Straw in the wind test. This test may question the hypothesis slightly or increase its plausibility marginally. It is *neither necessary nor sufficient* and is the weakest of the four tests. As such, it also places the “least demand on the researcher’s knowledge and assumptions” (Collier 2010). In the legal accountability example, more or less timely prosecution of the recommended cases referred to in the State Audit Report can be observed. Either way, this fact makes a less significant contribution to the conclusion about the state of legal accountability.

Doubly decisive test. This test is “a necessary and sufficient criterion for accepting a hypothesis” (Bennett 2010). In practice, the combination of a *hoop* test and a *smoking gun* test can achieve the same goal as a doubly-decisive test, since it is rare in social science for a single piece of evidence to affirm a hypothesis. For example, if evaluators observe that courts are prosecuting the majority of important cases referred to in the State Audit Report (hoop test), the practice of intransparent bidding is decreasing (smoking gun test), the courts are conducting their hearings and delivering their judgments in a fair, effective and efficient way and public trust is growing (smoking gun test), they can then combine all of these tests to form a doubly decisive test and conclude that the prosecution of cases referred to in the State Audit Report has led to positive changes in legal accountability.

Remarks on process-tracing

Strong internal validity. Process-tracing, being a prominent tool of within-case analysis, can help evaluators to draw causal inferences that have strong internal validity. Internal validity refers to “*the validity of inferences about whether observed co-variation between A (the presumed treatment) and B (the presumed outcome) reflects a causal relationship from A to B as those variables were manipulated or measured*” (Shadish / Cook / Campell 2002, 34). In plain English, when we talk about internal validity, we are interested in whether A causes B. Three conditions must hold if we are to draw sound conclusions about this relationship: (i) A must precede B in time, (ii) A co-varies with B, and (iii) no other explanation for the relationship is plausible (Shadish / Cook / Campell 2002, 53).

Process-tracing satisfies the first condition for internal validity in that it establishes temporal sequences of causal events and examines the process of change (Munch 2004). The method fulfils the third condition through its capacity to deal with selection bias, one of the major threats to internal validity. The threat here is that intrinsic differences in the unit of analysis, rather than the intervention/treatment, cause the observed effects. The quantitative approach tries to solve this problem through randomisation and experiment, which it is hoped will distribute the threat randomly and so equalise characteristics among control and treatment groups (Shadish / Cook / Campell 2002, 249). In process-tracing, proximity to the case allows evaluators to examine in depth each of the key explanatory factors and their relative importance in producing effects. More importantly, evaluators can verify the role of plausible alternative explanatory factors by studying the process and mechanism through which they insert changes.

Weak on generalisation. The strength of the method, however, is also its weakness. Nuanced understanding of the case comes at the cost of generalisation, i.e. each case is a unique configuration of context, explanatory factors, actors, mechanism and causal process. While this approach can generate better internal validity, discover important missing variables and unearth unintended effects, its findings are essentially idiosyncratic to the case. This has direct implications for the feasibility of evaluating the effectiveness and governance effects of budget support. The method is demanding on evaluators’ background knowledge of the subject and understanding of the local political and institutional set-up. The proposed solution is that evaluators develop sufficient understanding of all cases and categorise cases according to causal mechanism, policy fields, level of targeted policy changes or other important criteria. As cases within a category are similar in critical causal respects and so permit meaningful generalisation, evaluators can select one or two typical cases to conduct in-depth process-tracing.

Lack of guidance during the observation process. Scholars have taken a long step towards providing a clear set of logic for causal inference. However, the method still lacks guidance and procedures for evaluators/researchers *during* the data collection period. By this I mean guidance on *what* to observe, not *how* to observe. The literature on qualitative research and evaluation methods is full of guidance on how to observe. For instance, the observation method prescribes different dimensions of fieldwork observations, depending on the purpose of the study. Evaluators can decide to engage in short, one-shot or long, multiple

observations. They can also choose different degrees of immersion, and the intensity alters with the change from onlooker to participant status. Advice is available on how to integrate key informants and observe what did not happen during the inquiry, etc. (Patton 2002).

What is missing here is a step-by-step guide that helps evaluators to map their way from inputs, context, through causal steps and elements of causal mechanism to the stage where outcomes and effects emerge. Although the analytical framework plays a critical role in this respect, there is a limit to evaluators' prior knowledge. Moreover, more than one causal mechanism may be at work, actors may interact in unexpected ways owing to new developments in their environments, and unintended effects may parallel intended ones. In short, the guide is needed to enable the causal process to be systematically discovered and captured on the basis of limited prior knowledge. For example, historical process-tracing (Clayton 1996) categorises different types of event structure and provides practical advice on how to trace out the causes for each type of event. The guide tells a researcher which path to follow or ignore, how to distinguish a crucial cause from a trivial one and where to start, stop or make a detour in the tracing process. Political studies need something equivalent.

3.2 Evaluating political conditionality and policy dialogue

What is political conditionality?

Political conditionality is another aspect of budget support conditionality. It serves the democratic governance objective of budget support. The idea here is that budget support can be used as an instrument to leverage good governance. This objective contrasts with the technocratic objective of budget support, i.e. it can be used as leverage for reforms and thus better poverty reduction records (Molenaers / Cepinkas / Jacobs 2010). The democratic governance objective is quite explicit for some budget support donors, such as Germany:

“Governance objective: Budget support programmes contribute to reform processes in the partner country that aim to building functioning public institutions on a sustainable basis and, in particular, improve transparency, accountability, and the effectiveness and efficiency of public administration and public financial management. Budget funding helps to improve specialised and democratic control of expenditure and policy objectives by improving accountability to parliament and civil society. Budget support programmes also promote political dialogue on respect for and realisation of human rights, democratic participation, the rule of law and gender equality.” (BMZ 2008, 8)

While technical conditionality largely forms part of the Performance Assessment Framework, political governance conditionality can largely be found in the Underlying Principles (UP) agreed between BS donors and the partner government. An examination of the Underlying Principles of six countries (Ghana, Malawi, Mozambique, Rwanda, Zambia and Tanzania) reveals five common elements. Essentially, the UP require commitments from partner governments in five areas: (i) to pursue sound macroeconomic policies and management, (ii) to commit to poverty reduction and the achievement of the Millennium Development Goals through the implementation of the national development strategy, (iii)

to maintain sound budgeting and public financial systems and to commit to reforms in the respective fields, (iv) to safeguard peace, human rights, the rule of law, democratic principles and an independent judiciary and (v) to adhere to good governance principles and to fight corruption.

It is important to note in the case of the political conditions that the line between technical and political is at best fuzzy. The UP list comprises issues that are both of a purely democratic nature (human rights, rule of law) and of a more technocratic nature (macroeconomic management, poverty reduction). The integration of the two into one set of UPs reflects the “profound political” implications of technical issues:

“Politics is defined as all those activities of cooperation, conflict, bargaining over the production, allocation and distribution of tangible and intangible resources (Leftwich 1996). Politics is about power, about who gets what, when and how. Politics always entails preferences, it’s about making choices (Heyden 2005). From that perspective, technocratic governance is profoundly political, because when it comes to designing and implementing macroeconomic policies, composing the budget, dealing with sector reforms, public finance management, etc there are always choices to be made, between policies, between priorities, between goals and objectives.” (Molenaers / Cepinkas / Jacobs 2010, 7)

BS political conditionality has two main characteristics: it is closely associated with leverage, and yet it is vague. In the case of technical conditionality, BS donors go to great lengths to craft sophisticated disbursement mechanisms to reflect the actual level of performance and increase aid predictability. However, the rules of the game change fundamentally when it comes to political conditionality. Violation of the core elements of the UPs may, for example, lead to *“immediate and complete cessation of cooperation on budget support if political dialogue fails to produce viable solutions”* (BMZ 2008, 18). In practice, a study of five donors by Molenaers / Cepinkas / Jacobs (2010, 14) reveals that donors react to a breaching of the UPs in four ways:

- (i) the suspension of the entire aid envelope, including BS,
- (ii) the suspension of BS and the channelling of the funds to actors other than the government,
- (iii) a reduction in BS and sometimes the whole aid envelope and
- (iv) a delay in BS and/or the entire aid envelope.

Secondly, political conditions are vague: they are expressed in general terms and are not specified in the Underlying Principles. In technical conditionality, the financial reward depends on a clear list of policy actions to be taken and targets to be achieved by partner governments. In the case of political conditions, financial rewards are linked to actions that partner governments should not take. As those actions cannot be foreseen, BS donors observe the governments’ performance and react when they deem a breach has occurred. According to Molenaers / Cepinkas / Jacobs (2010), the advantage of this approach is that BS donors have the leeway to respond, but the drawback is that it is reactive rather than proactive. In other words, BS donors employ political conditionality as a stick when things go wrong, but not strategically as a carrot (ibid).

For an evaluator evaluating the effectiveness of political conditionality means asking whether BS donors achieve their objective with this particular type of conditionality. Or, to quote Hayman, who studies the impact of BS political conditionality:

“That is, does the threat of or the actual withholding of budget support sway recipient governments into a different course of action?” (2011, 638)

“Summative evaluations serve the purpose of rendering an overall judgment about the effectiveness of a program, policy, or product for the purpose of saying that the evaluand (thing being evaluated) is or is not effective and, therefore, should or should not be continued, and has or does not have the potential of being generalizable to other situations” (Patton 2002, 218)

Why evaluate political conditionality and policy dialogue together?

In fact, BS conditionality and policy dialogue are so intertwined that evaluators rarely distinguish the two components. Any evaluation of BS conditionality inevitably includes the work of policy dialogue, since BS donors and their partner governments agree on the terms and conditions of conditionality through policy dialogue. Similarly, any evaluation of policy dialogue inherently covers conditionality, since the core elements of conditionality are the financial rewards, threats of sanctions, policy priorities and other preferences of the parties concerned that shape the outcomes of policy dialogue.

The ambiguous and ad-hoc nature of political conditionality magnifies the role of policy dialogue. In the absence of a programme document prescribing a logical framework or a theory of change, BS donors and partner governments agree on the most basic elements of political conditionality in a series of policy dialogues. For example, BS donors need first to establish that a breach of the Underlying Principles has occurred and agree on the solutions, including the desirable course of action, the concrete threat of sanctions, etc. The negotiations take place not only between BS donors and partner governments, but also among BS donors. The same study by Molenaers / Cepinkas / Jacobs (2010) notes different perceptions of a “breach” by BS donors, which leads to “substantial differences in ‘identifying’ or ‘labelling’ an event as problematic or a crisis” (ibid., 13):

- *“A breach is a fundamental and extreme reversal of the political system/situation, like a coup.*
- *A breach is a deterioration of the Ups*
- *A breach is when there is no progress on Ups”*

Taken together, evaluators investigating the effectiveness of BS political conditionality must examine the policy dialogue process. Consequently, it is futile to separate the two components for evaluation purposes. The next section proposes a possible design for evaluating the effectiveness of BS political BS conditionality (and thus policy dialogue).

A possible design

Choice of method: the process as a nature of intervention and process evaluation

If we view political conditionality as an intervention to be evaluated, one important characteristic of this intervention is that it is highly process-oriented. Typically, as the first step, evaluators will verify that a program logical framework and its critical elements are in place. The outcomes should be clearly defined and ready to be evaluated, the intervention should be supported by a plausible, identifiable theory of change that logically connects implementation with outcomes (Patton 2002). In BS political conditionality, this programme logic is absent and only gradually emerges once the crisis starts. For example, once the partner government has decided to use public money to save badly managed, semi-private banks, or once military expenditure has surged, or a corruption scandal has erupted, BS donors formulate the desirable outcomes, often a reverse of the perceived breach of the Underlying Principles. Starting from the outcomes, BS donors first individually and then collectively devise the intervention which it is hoped will remedy the situation. Various means are considered, and the real leverage very often continues to be the credible threat of financial sanction. The dynamic of this formulation-via-negotiation process therefore has direct implications for the outcome, i.e. whether BS donors are successful in “swaying” the partner government to take what they consider the desirable course of action.

The most suitable type of evaluation for this case would be *process evaluation*. The decisive impact of the process on the outcome largely justifies this recommendation. Secondly, evaluators do not have a blueprint or programme document to guide their measurement. The explanatory factors and the underlying intervention logic are not written *ex ante*. However, evaluators are able to grasp the most critical elements of this intervention logic as they play out during the negotiation process. Evaluators can thus capture the explanatory factors and the outcome by examining the entire process of the crisis. Patton (2002) has summarised the goal of process evaluation as follows:

“Process evaluations aim at elucidating and understanding the internal dynamics of how a program, organization, or relationship operates A process evaluation not only looks at formal activities and anticipated outcomes, but they also investigate informal patterns and unanticipated interactions...By describing and understanding the details and dynamics of program processes, it is possible to isolate critical elements that have contributed to program successes and failures” (Patton 2002, 159–160).

Data collection

Evaluators should first select the *unit of analysis*, i.e. the individuals, organisations, entities or objects on which they will collect data (Brady / Collier 2004, 311). The unit of analysis matters because it determines the inference and conclusion that an evaluator may draw at the end of the data analysis process. As the outcome of BS political conditionality consists of a series of policy dialogues which shape the final outcome, it is suggested that evaluators focus on each dialogue as a unit of analysis. It is necessary for evaluators to break down the important elements of policy dialogues and collect all the concerned quantitative and qualitative data. For instance, the actors, their motivation, the actors' accountability structure, the agenda, the dynamic of the negotiation and the dialogue outcomes are essential factors in policy dialogue.

The second element of a research/evaluation design is *sampling*, i.e. choosing which unit of analysis to observe. Qualitative methods are characterised by purposeful sampling, i.e. choosing cases because they are information-rich and may provide useful insights and an in-depth understanding (Patton 2002, 230). From the wealth of literature on purposeful sampling for the qualitative method, evaluators should select the sampling strategy that may best reveal how the policy dialogue process leads to an outcome. For example, if BS donors held a dozen dialogues to diagnose a situation and agree on the solutions, it would not be practical to study all of them. Evaluators can then use *extreme or a deviant-case sampling* strategy for in-depth examination: "*this strategy involves selecting cases that are information rich because they are unusual or special in some way, such as outstanding success or notable failures*" (Patton 2002, 232). Evaluators might therefore select dialogues in which BS donors fail to agree on any of the important aspects, e.g. a desirable course of action for the government, sanctions for non-compliance, etc. Studies of such dialogues would reveal the underlying factors that are obstructing efficient collaboration. Alternatively, evaluators could focus on policy dialogues in which the BS donor group manages to agree on notoriously difficult issues. For instance, as BS donors all have different priorities when it comes to policy areas, they adopt different disbursement mechanisms. Yet they may be unanimous in choosing the type of financial sanction in the event of non-compliance. Studies of such dialogues would throw light on a few key elements that are crucial for all BS donors and might therefore make for more efficient collaboration.

Another attractive sampling strategy is *maximum variation (heterogeneity) sampling*. In this strategy, evaluators aim at "*capturing and describing the central themes that cut across a great deal of variation*" (Patton 2002, 235). Evaluators should first establish criteria for classifying and selecting cases. For example, a high-level dialogue versus a dialogue among the rank and file of the two sides will be an important dimension for case selection. Alternatively, evaluators may select dialogues representing different *stages* of the negotiation process. This sampling strategy is attractive in that it can help evaluators to extract the common elements that persistently stand out of the variation and could thus help to explain the outcome.

Data analysis

Box 6: Options for organizing and reporting qualitative data	
Story-telling approaches	
Chronology and History	Describe what happened chronologically over time, from beginning to end
Flashback	Start from the end, work backwards to describe how the ending emerged
Case study approaches	
People	If individuals or groups are the primary unit of analysis, then case studies of people or groups may be the focus for case studies
Critical incidents	If critical incidents or major events are the unit of analysis, description follows order of importance rather sequence of occurrence
Various settings	Describe various places, sites, settings or locations before doing cross-setting pattern analysis
Analytical framework approaches	
Processes	Organise data into important processes, and distinguishing important processes becomes the analytical framework for organising qualitative description
Issues	An analysis can be organised to illuminate key issues, often the equivalent of the primary evaluation questions
Questions	Response to interviews can be organised question by question, especially where a standardised interviewing format is used
Sensitising concepts	When sensitising concepts have played an important role in guiding fieldwork, the data can be organised and described through sensitising concepts
Source: reproduced from Patton (2002, 439)	

In *data analysis*, the purpose of the evaluation should guide the data analysis. Typically, data analysis consists of two stages: description and drawing inferences. At the end of the data analysis process it should be clear (i) whether BS political conditionality leads to positive changes in specific aspects of the partner government's political governance and (ii) which factors contribute to success and failure. Data analysis methods should therefore be chosen in a way that best answers these questions.

At the first stage of description, evaluators have the choice of adopting a story-telling, case-study or analytical framework approach to organise their data. Each approach has further choices, the storytelling approach, for example, consisting of chronology and flashback approaches. The themes and pattern of the data should determine the way those data are reported. As we are conducting process evaluation, it is natural to resort to negotiation processes for the reporting of data. However, evaluators can also adopt other approaches to

summarise data and information if they find these formats capture the nature of the data and findings better. If, for example, the evaluators detect a series of issues (harmonisation among BS donors, leverage as an effective means, negotiating skills, etc.) that defines the dialogue outcomes, then data should be reported according to issues. In the study of BS political conditionality, Molenaers / Cepinkas / Jacobs (2010) describe their data and information in *chronological* order, i.e. before, during and after the crisis. This way of reporting enables the authors to highlight one important finding: BS donors became progressively more frustrated over a long period of time with respect to governance issues, the eventual eruption being caused by one single event. Apart from the conclusion that harmonisation among BS donors, which lends credibility to threats, explains the effectiveness of political conditionality, the authors therefore identify a major deficiency in the design of BS conditionality and come up with the following recommendation:

“One of the biggest advantages of result-oriented aid, is at the same time its biggest disadvantage. By pinning everything down in indicators, the legitimate scope of discussion is narrowed to that framework (in this case the PAF). Problems or concerns that pop up in other areas, or concerns that go beyond one specific reform, may become ‘untouchable’ The first relevant question that should be posed is how to organize for a fora in such a way that growing discontent/frustration can be channelled and voiced.”
(Molenaers / Cepinkas / Jacobs 2010, 44–45)

In summary, as evaluators transform raw data into meaningful findings, they have various options as regards how to report and what to describe. The guiding principle for this process is to select the most effective framework for communicating the data. That, in turn, means giving preference to the voice of the data, i.e. detecting significant patterns in order to capture the essence of the data (Patton 2002, 432). In BS political conditionality, evaluators identify from the raw data events, issues, processes, incidents or factors that critically define the outcome and report according to these dimensions.

The second and also the last stage of data analysis is the drawing of *causal inferences*. Either of two “tools” can accomplish this task: counterfactual analysis or causal mechanisms. The section on process-tracing introduced the four tests for guiding causal inference. These tests do not apply here because, in the absence of an elaborated logical framework and strong assumptions about how BS political conditionality is meant to work, it is impossible to formulate ex-ante hypotheses for testing. However, counterfactual analysis and causal mechanisms are two powerful means of drawing causal conclusions for qualitative researchers.

The quantitative approach uses experiments to estimate the effect of the treatment without the causal mechanism being observed. The qualitative approach, on the other hand, makes a causal inference by trying to fill in this “*black box*”, i.e. by establishing the causal mechanism and the pathway from intervention to effect (Goertz / Mahoney s. a). The causal mechanism therefore becomes an important means of drawing causal inferences:

“They [the researchers] see mechanisms as a nonexperimental way of distinguishing causal relations from spurious correlations: Mechanisms help in causal inference in two ways. The knowledge that there is a mechanism through which X influences Y supports the inference that X is a cause of Y. In addition, the absence of a plausible mechanism

linking X to Y gives us a good reason to be suspicious of the relation being a causal one ...” (ibid., 99).

If evaluators demonstrate convincingly through their description that a causal mechanism explains how political conditionality exercises its influence and changes the partner government’s course of action, that description is sufficient for a causal inference to be drawn. For example, the data may give a clear picture of the key causal agent, whether one individual, collective actors, and of the evidence of the type of causal mechanism at work, e.g. rational choice, power reproduction or positive feedback. When such ideas about the causal pathway and mechanism emerge clearly from the data, evaluators can draw valid causal conclusions. Having studied thousands of scholarly works and the theory and practice of qualitative research and evaluation, Patton (2002, 479) provides guidance for this step:

“...Once case studies have been written and descriptive typologies have been developed and supported, the task of organization and description are largely complete and it is appropriate, if desired, to move on to make comparisons and consider causes, consequences and relationships When careful study of the data gives rise to ideas about causal linkages, there is no reason to deny those interested in the study’s results the benefit of those insights.”

The second option is to conduct counterfactual analysis to assess the role of BS political conditionality in changes to the partner government’s course of action. Qualitative researchers commonly use counterfactual analysis to solve the fundamental problem of causal inference. The essence of the problem stems from the fact that the effects of treatment and non-treatment on the same unit cannot be observed at the same time (Holland 1986, 974). Qualitative researchers employ within-case analysis to construct a claim of what would have happened in a particular case, for a specific outcome. The reasoning will typically be based on a sequence of events without intervention (called X). If these sequences of reasoning do not lead to the outcome (Y), it can be inferred that X is an important cause of Y. The use of counterfactual analysis is demanding on researchers’/evaluators’ knowledge of a particular case if it is to produce a plausible counterfactual.

Several principles rule the construction of counterfactual analysis. According to the Minimum Rewrite Rule, the counterfactual should entail changing the historical facts and records as little as possible (Stalnaker 1986, 104; Elster 1978; Tetlock / Belkin 1996, 23–25; Reiss / 2009, quoted by Goertz / Mahoney s. a.). To comply with this principle, evaluators should avoid “miracle” counterfactual conditions, or a big “if”. In a similar vein, the outcome of the counterfactual reasoning should normally materialise in a real world of likewise situations (Lewis 1973, quoted by Goertz / Mahoney s. a.). The second rule concerns the length of the causal argument. Clayton (1996) suggested that counterfactual argument can be “safely made” only when it concerns but one step in the causal sequence, since each step combines several causal factors and they cannot all be known to the researcher. Moreover, chance plays a role in shaping the combination of causal factors, which can never be predicted ex post.

As an example, Howard White and Geske White Dijkstra (2003) conducted a series of case studies on structural adjustment conditionality and reforms. In the case of Nicaragua, the authors depicted the country's historical context and the macroeconomic environment of the reform. They provided descriptions, conducted an analysis from macro to meso level and demonstrated how political events and constellations of actors influence economic management and how losers and winners determine the path of reform. Finally, the assessment of aid impact was based on counterfactual arguments relating to different scenarios of what 'would have been'. This is a good example of how within-case analysis can be conducted using an analytical framework from macro to micro level, in combination with narrative analysis and simple counterfactual argument.

4 Conclusions and recommendations

This paper sets out to investigate the question of how to use qualitative methods to evaluate the effectiveness of BS conditionality and policy dialogue. BS conditionality consists of technical and political conditionality. Each has its own intervention logic and objectives. It is argued that the necessary evaluation framework has still to be constructed. The paper then proceeds to propose an evaluation framework for BS technical conditionality, political conditionality and policy dialogue. Each evaluation framework has two parts: evaluation questions with hypothesised causes and effects and an evaluation design to answer those questions.

In BS technical conditionality, the main questions are (i) whether the conditionality is complied with and which factors contribute to the observed performance, (ii) what are the effects of conditionality, once it is complied with and (iii) what are the governance effects of BS technical conditionality. Qualitative Comparative Analysis (QCA) can answer the first evaluation question. It identifies combinations of causes that produce the outcome by detecting and systemising patterns of causes and deriving the causal configuration from the data. The advantage of this method is that it permits generalisation, while retaining a detailed understanding of the case. It can also bring important missing variables to light and identify different causal pathways to the outcome. On the other hand, it is demanding on evaluators' prior knowledge of the causal factors and requires a laborious exercise of moving back and forth between data analysis and case investigation.

Process-tracing of causal mechanisms is the method selected to answer the next two evaluation questions concerning BS technical conditionality. It is able to identify changes and draw causal conclusions by providing evidence of a causal pathway leading from intervention to outcome. This method permits causal conclusions with strong internal validity, but it has the drawback of limited capacity for generalisation. It is also impractical when it comes to conducting good-quality process-tracing for more than a handful of cases, and more thought on case selection strategy to reduce the number of cases to be studied is called for.

In answer to the evaluation question relating to political conditionality and policy dialogue, this paper proposes the use of process evaluation in combination with counterfactual

analysis. Owing to the distinctive characteristics of political conditionality, its intervention logic plays out only in policy dialogue processes. The link between the two components makes it more useful to evaluate the two components together. Process evaluation helps to identify events, processes, issues, factors and other important dimensions that shape the outcome. Counterfactual analysis can be used to attribute the outcome to the intervention of BS political conditionality and policy dialogue or other factors. The method has proved to be effective and efficient in assisting scholars to study the effectiveness of structural adjustment conditionality and BS policy dialogue.

In conclusion, qualitative methods can provide answers to evaluation questions. Such tools as Qualitative Comparative Analysis, process-tracing and counterfactual analysis permit sound causal inferences, identify important missing variables and make observations on causal processes and mechanisms possible. These methods are thus able not only to answer the evaluation questions, i.e. assess the effectiveness of an intervention, but also to shed light on the processes and factors that limit or facilitate the success or failure of the intervention. However, a number of challenges need to be addressed if these methods are to be used effectively and efficiently:

- *An analytical framework* that identifies the evaluation questions, the underlying intervention logic, the hypothesised causes, the effects and, at times, the causal mechanisms has still to be constructed. This framework is a critical starting point that would greatly facilitate the work of evaluators in the field, since it brings to the fore objectives, assumptions and hypotheses on how BS conditionality and policy dialogue are meant to work. The absence of this framework leaves the evaluator at a loss as to what to measure and makes it difficult to judge evaluation results. It is therefore recommended that the next step be an agreement among BS donors, managers and evaluators on an analytical framework for BS conditionality and policy dialogue. Ideally, this framework should be complemented by method guidelines so as to form an evaluation framework.
- *Further work on case selection.* Qualitative method researchers and evaluators traditionally work with a small N, i.e. a small number of cases. The reason for their working on few cases is not that they do not have more cases to study, a common misconception, but that a small number of cases enables them to examine the complexity of each and to *capture* rich and useful insights. There is, then, an inherent trade-off between depth and breadth when it comes to case selection. In evaluating the effectiveness of BS technical conditionality and its effects on governance, evaluators may face 30–100 individual conditions with idiosyncratic intervention logic. It is virtually impossible to carry out good quality process-tracing on all of them. This paper makes a number of recommendations for remedying the situation. Evaluators can first focus on *positive cases*, i.e. cases where the compliance level is high. Considering whether the objective of conditionality has been achieved is futile if the condition itself is not fulfilled in the first place. The second strategy is to categorise cases according to important criteria and to select a few for intensive study, so that the result of these cases may be meaningfully generalised to the entire category. One such criterion might be the type of causal mechanism. As the next step, evaluators, in consultation with BS donors and managers, might develop sets of criteria for categorising technical conditionality for evaluation purposes.

- *Process-tracing technique.* Scholars of qualitative methods have taken process-tracing a step forward by proposing tests with clear criteria for drawing causal inferences. A previously defined analytical framework also helps evaluators considerably with data collection. *However*, there is a “missing middle” in methodological guidance: how to manoeuvre and trace steps of the causal pathway when there are discrepancies between the framework and reality. Political studies need something equivalent to the guidance provided on historical process-tracing, which tells researchers how to differentiate a crucial and a trivial cause, when they have come to a dead end and need to make a detour, etc. It is recommended that future attempts to improve the process-tracing technique take up the issue of how to conduct process-tracing in different types of political processes. What is needed here is guidance on the systematic identification and capture of causal processes when prior knowledge is limited.

Bibliography

- Bennett, A. (2004): Case study methods: design, use and comparative advantages, in: D. F. Sprinz / Y. Wolinsky-Nahmias (eds.), *Models, numbers, and cases: methods for studying international relations*, Ann Arbor, Mich.: University of Michigan Press, 19–55
- (2010): Process tracing and causal inference, in: E. H. Brady / E. Henry / D. Collier: *Rethinking social inquiry: diverse tools, shared standards*, Lanham, Md.: Rowman & Littlefield, 2nd ed., 207–220
- Brady H. E. / D. Collier (2004): *Rethinking social inquiry*, Lanham, Md.: Rowman & Littlefield
- Drazen, A. / P. Isard (2004): *Can public discussion enhance program ownership?*, Cambridge, Mass.: National Bureau of Economic Research (Working Paper 10927)
- BMZ (*Bundesministerium für wirtschaftliche Zusammenarbeit und Entwicklung*) (2008): *Budget support in the framework of programme-oriented joint financing*; online: <http://www.bmz.de/en/publications/topics/financing/konzept181.pdf> (accessed 6 Mar. 2012)
- Brinkerhoff, D. W. (1996): Process perspective on policy change: highlighting implementation, in: *World Development* 24 (9), 1395–1401
- Caputo, E. / A. Lawson / M. van der Linde (2008): *Methodology for evaluations of budget support operations at country level*; online: http://www.worldbank.org/ieg/nonie/docs/issue_paper.pdf (accessed 6 Mar. 2012)
- Clayton, R. (1996): *The logic of historical explanation*, University Park, Pa.: Penn State University Press
- Collier, D. (2010): *Process tracing: introduction and exercise*, beta version, manuscript
- Compernelle, P. / A. de Kemp (2009): *Methodology for evaluation of budget support operations at country level*, mimeo
- DFID (*Department for International Development*) (2005): *Partnerships for poverty reduction: rethinking conditionality*; online: <http://www2.ohchr.org/english/issues/development/docs/conditionality.pdf> (accessed 6 Mar. 2012)
- Dijkstra, G. (2011): The PRSP approach and the illusion of improved aid effectiveness: lessons from Bolivia, Honduras and Nicaragua, in: *Development Policy Review* 29 (1), 111–133
- Dom, C. (2007): *What are the effects of general budget support?*, Birmingham: International Development Department of the University of Birmingham
- Dreher, A. (2009): IMF conditionality: theory and evidence, in: *Public Choice* 141 (1), 233–267
- Falletti, T. G. / J. F. Lynch (2009): Context and causal mechanisms in political analysis, in: *Comparative Political Studies* 42 (9), 1143–1166
- Faust, J. (2010): Policy experiments, democratic ownership and development assistance, in: *Development Policy Review* 28 (5), 515–534
- Faust, J. / S. Leiderer / S. Koch (2011): *Multi-donor budget support: only halfway to effective coordination*, Bonn: DIE (Briefing Paper 8/2011)
- Gerring, J. (2008): Case selection for case-study analysis: qualitative and quantitative techniques, in: J. M. Box-Steffensmeier / E. H. Brady / D. Collier: *The Oxford Handbook of political methodology*, Oxford: Oxford University Press, 645
- Goertz, G. (2006): *Social science concepts: a user's guide*, Princeton, NJ: Princeton University Press
- Goertz, G. / J. Mahoney (s.a.): *Causal models*, in: *A tale of two cultures: contrasting qualitative and quantitative paradigms*, Princeton, NJ.: Princeton University Press
- Gunning, J. W. / C. Elbers / K. de Koop (2009): Assessing sector-wide programs with statistical impact evaluation: a methodological proposal, in: *World Development* 37 (2), 513–520

- Hayman, R. (2011): Budget support and democracy: a twist in the conditionality tale, in: *Third World Quarterly* 32 (4), 673–688
- Holland, P. W. (1986): Statistics and causal inference, in: *Journal of the American Statistical Association* 81 (396), 945–960
- IEG (*Independent Evaluation Group*) (2009): Poverty reduction support credits: an Evaluation of World Bank support 2009, Washington, DC
- (2010): Poverty reduction support credits: an evaluation of World Bank support 2010, Washington, DC
- Kemp, A. de / J. Faust / S. Leiderer (2011): Between high expectations and reality: an evaluation of budget support in Zambia (2005–2010: synthesis report; online: <http://www.oecd.org/dataoecd/25/31/49210553.pdf> (accessed 6 Mar. 2012)
- Killick, T. (1997): Principals, agents and the failings of conditionality, in: *Journal of International Development* 9 (4), 483–495
- Killick, T. / R. Gunatilaka / A. Marr (1998): Aid and the political economy of policy change, London: Routledge
- King, G. / R. O. Keohane / S. Verba (1994): Designing social inquiry: scientific inference in qualitative research, Princeton, NJ: Princeton University Press
- Koeberle, S. / Z. Stavreski (2006): Budget support: concept and issues, in: S. Koeberle / Z. Stavreski / J. Walliser: Budget support as more effective aid?: recent experiences and emerging lessons, Washington, DC: World Bank Publications, 3–23
- Love, A. (2004): Implementation evaluation, in: J. S. Wholey / H. P. Hatry / K. E. Newcomer, Handbook of practical program evaluation, Hoboken, NJ: John Wiley & Sons, 2nd ed., 63–98
- Mahoney, J. (2000): Strategies of causal inference in small-n analysis, in: *Sociological Methods & Research* 28 (4), 387–424
- Maxwell, A. J. (2004): Using qualitative methods for causal explanation, in: *Field Methods* 16 (3), 243–264
- McLaughlin J. A. / G. B. Jordan (2004): Using logic models, in: J. S. Wholey / H. P. Hatry / K. E. Newcomer, Handbook of practical program evaluation, Hoboken, NJ: John Wiley & Sons, 2nd ed., 7–33
- Mill, J. S. (1882): A system of logic, ratiocinative and inductive: being a connected view of the principles of evidence and the methods of scientific investigation, New York City, NY: Harper & Brothers
- Molenaers, N. / L. Cepinskas / B. Jacobs (2010): Budget support and policy, political dialogue: donor practices in handling (political) crises; online: <http://www.ua.ac.be/objs/00256641.pdf> (accessed 6 Mar. 2012)
- Morrissey O. / A. Verschoor (2004): What does ownership mean in practice?: policy learning and the evolution of pro-poor policies in Uganda: paper presented at the HWWA conference for the political economy of aid; online: <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.201.6299&rep=rep1&type=pdf> (accessed 6 Mar. 2012)
- North, D. C. (1991): Institutions, institutional change and economic performance, Cambridge: Cambridge University Press
- OECD (*Organisation for Economic Co-operation and Development*) (2006): Harmonising donor practices for effective aid delivery, vol. 2, Paris
- (2011): 2011 Survey on monitoring the Paris Declaration; online: <http://www.oecd.org/dac/pdsurvey> (accessed 6 Mar. 2012)
- Patton, M. Q. (2002): Qualitative research and evaluation methods, Thousand Oaks, Calif.: Sage
- Przeworski, A. / H. Teune (1970): The logic of comparative social inquiry, Malabar, Fla. Krieger Publishing
- Ragin, C. C. (2000): Fuzzy-set social science, Chicago, Ill.: University of Chicago Press

- Rihoux, B.* (2006): Qualitative comparative analysis and related systematic comparative methods: recent advances and remaining challenges for social science research, in: *International Sociology* 21 (5), 679–706
- Rihoux, B. / C. C. Ragin* (2009): Configurational comparative methods: qualitative comparative analysis and related techniques, Thousand Oaks, Calif.: Sage
- Schmidt, P.* (2006): Budget support in the EC's development cooperations, Bonn: Deutsches Institut für Entwicklungspolitik (Studies 20)
- Shadish, W. R. / T. D. Cook / D. T. Campbell* (2002): Experimental and quasi-experimental designs for generalized causal inference, Boston, Mass.: Houghton Mifflin
- Shand, D.* (2006): Managing fiduciary issues in budget support operations, in: S. Koeberle / Z. Stavreski / J. Walliser: Budget support as more effective aid?: recent experiences and emerging lessons, Washington, DC: World Bank Publications, 28–40
- Stokke, O. S.* (2007): Qualitative comparative analysis, shaming, and international regime effectiveness, in: *Journal of Business Research* 60 (5), 501–511
- White, H. / G. Dijkstra* (2003): Programme aid and development: beyond conditionality, London: Routledge, Taylo & Francis Group
- World Bank* (2006): Good practice principles for the application of conditionality: a progress report, Washington, DC: World Bank Publications

Publications of the German Development Institute

Nomos Verlagsgesellschaft

- Messner, Dirk / Imme Scholz* (eds.): *Zukunftsfragen der Entwicklungspolitik*, 410 p., Nomos, Baden-Baden 2004, ISBN 3-8329-1005-0
- Neubert, Susanne / Waltina Scheumann / Annette van Edig, / Walter Huppert* (eds.): *Integriertes Wasserressourcen-Management (IWRM): Ein Konzept in die Praxis überführen*, 314 p., Nomos, Baden-Baden 2004, ISBN 3-8329-1111-1
- Brandt, Hartmut / Uwe Otzen*: *Armutorientierte landwirtschaftliche und ländliche Entwicklung*, 342 p., Nomos, Baden-Baden 2004, ISBN 3-8329-0555-3
- Liebig, Klaus*: *Internationale Regulierung geistiger Eigentumsrechte und Wissenserwerb in Entwicklungsländern: Eine ökonomische Analyse*, 233 p., Nomos, Baden-Baden 2007, ISBN 978-3-8329-2379-2 (Entwicklungstheorie und Entwicklungspolitik 1)
- Schlumberger, Oliver*: *Autoritarismus in der arabischen Welt: Ursachen, Trends und internationale Demokratieförderung*, 225 p., Nomos, Baden-Baden 2008, ISBN 978-3-8329-3114-8 (Entwicklungstheorie und Entwicklungspolitik 2)
- Qualmann, Regine*: *South Africa's Reintegration into World and Regional Markets: Trade Liberalization and Emerging Patterns of Specialization in the Post-Apartheid Era*, 206 p., Nomos, Baden-Baden 2008, ISBN 978-3-8329-2995-4 (Entwicklungstheorie und Entwicklungspolitik 3)
- Loewe, Markus*: *Soziale Sicherung, informeller Sektor und das Potenzial von Kleinstversicherungen*, 221 p., Nomos, Baden-Baden 2009, ISBN 978-3-8329-4017-1 (Entwicklungstheorie und Entwicklungspolitik 4)
- Loewe, Markus*: *Soziale Sicherung in den arabischen Ländern: Determinanten, Defizite und Strategien für den informellen Sektor*, 286 p., Nomos, Baden-Baden 2010, ISBN 978-3-8329-5586-1 (Entwicklungstheorie und Entwicklungspolitik 7)
- Faust, Jörg / Susanne Neubert* (Hrsg.): *Wirksamere Entwicklungspolitik: Befunde, Reformen, Instrumente*, 432 p., Nomos, Baden-Baden 2010, ISBN 978-3-8329-5587-8 (Entwicklungstheorie und Entwicklungspolitik 8)

[Books may be ordered only through publishing house or bookshops.]

Book Series with Routledge

- Brandt, Hartmut / Uwe Otzen*: *Poverty Orientated Agricultural and Rural Development*, 342 p., Routledge, London 2007, ISBN 978-0-415-36853-7 (Studies in Development and Society 12)
- Krause, Matthias*: *The Political Economy of Water and Sanitation*, 282 p., Routledge, London 2009, ISBN 978-0-415-99489-7 (Studies in Development and Society 20)

[Books may be ordered only through publishing house or bookshops.]

Springer-Verlag

- Scheumann, Waltina / Susanne Neubert / Martin Kipping* (eds.): *Water Politics and Development Cooperation: Local Power Plays and Global Governance*, 416 p., Berlin 2008, ISBN 978-3-540-76706-0

Studies

- 64 *Ashoff, Guido et al.*: Evaluación del „Fondo de planificación estratégica e implementación de reformas autofinanciadas en Chile”, 92 p., Bonn 2012, ISBN 978-3-88985-501-5
- 63 *Ashoff, Guido et al.*: Evaluierung des deutsch-chilenischen “Fonds zur strategischen Planung und Umsetzung eigenfinanzierter Reformen”, 94 p., Bonn 2012, ISBN 978-3-88985-500-8
- 62 *Fues, Thomas / LIU Youfa*: Global Governance and Building a Harmonious World: A comparison of European and Chinese concepts for international affairs, 215 p., Bonn 2011, ISBN 978-3-88985-499-5
- 61 *Weikert, Jochen*: Re-defining ‘Good Business’ in the Face of Asian Drivers of Global Change: China and the Global Corporate Social Responsibility Discussion, 378 p., Bonn 2011, ISBN 978-3-88985-497-1
- 60 *Hampel-Milagrosa, Aimée*: The Role of Regulation, Tradition and Gender in Doing Business: Case study and survey report on a two-year project in Ghana, 77 p., Bonn 2011, ISBN 978-3-88985-496-4
- 59 *Weinlich, Silke*: Reforming Development Cooperation at the United Nations: An analysis of policy position and actions of major states on reform options, 134 p., Bonn 2011, ISBN 978-3-88985-495-7
- 58 *Chahoud, Tatjana et al.*: Corporate Social Responsibility (CSR) and Black Economic Empowerment (BEE) in South Africa: A case study of German Transnational Corporations, 100 p., Bonn 2011, ISBN 978-3-88985-494-0
- 57 *Neubert, Susanne et al.*: Agricultural Development in a Changing Climate in Zambia: Increasing resilience to climate change and economic shocks in crop production, 244 p., Bonn 2011, ISBN 978-3-88985-493-3
- 56 *Grimm, Sven et al.*: Coordinating China and DAC Development Partners: Challenges to the aid architecture in Rwanda, 200 p., Bonn 2010, ISBN 978-3-88985-492-6

[Price: 10,00 Euro; books may be ordered directly from the DIE or through bookshops.]

Discussion Paper

- 7/2012 *Faust, Jörg / Sebastian Ziaja*: German Aid Allocation and Partner Country Selection: Development-Orientatio´n, Self-Interests and Path Dependency, 27 p., Bonn 2012, ISBN 978-3-88985-551-0
- 6/2012 *Hensengerth, Oliver / Ines Dombrowsky / Waltina Scheumann*: Benefit-Sharing in Dam Projects on Shared Rivers, 37 p., Bonn 2012, ISBN 978-3-88985-549-7
- 5/2012 *Faust, Jörg*: Poverty, Politics and Local Diffusion: Resource allocation in Bolivia’s Decentralised Social Fund, 23 p., Bonn 2012, ISBN 978-3-88985-550-3
- 4/2012 *Marino, Robert / Ulrich Volz*: A Critical Review of the IMF’s Tools for Crisis Prevention, 38 p., Bonn 2012, ISBN 978-3-88985-547-3
- 3/2012 *Grävingholt, Jörn / Sebastian Ziaja / Merle Kreibaum*: State Fragility: Towards a Multi-Dimensional Empirical Typology, 38 p., Bonn 2012, ISBN 978-3-88985-546-6

[Price: 6,00 Euro; books may be ordered directly from the DIE or through bookshops.]

A complete list of publications available from DIE can be found at:
<http://www.die-gdi.de>