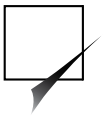

Article

Evaluation
2015, Vol. 21(4) 429–447

DOI: 10.1177/1356389015607739
evi.sagepub.com

The contribution of process tracing to theory-based evaluations of complex aid instruments

Johannes Schmitt
University of Duisburg-Essen; DEval – German Institute for Development Evaluation, Germany

Derek Beach
University of Aarhus, Denmark

Abstract
This article focuses on methodological challenges in evaluating complex program aid interventions like budget support. We show that recent innovations in process-tracing methodology can help solve the identified challenges and increase the strength of causal inference made when using case studies in demanding settings. For the specific task of evaluating the governance effectiveness of budget support interventions, we developed a more fine-grained causal mechanism for a subset of the comprehensive program theory of budget support. Moreover, based on the informal use of Bayesian logic, we have elaborated on how to increase the conclusiveness of empirical evidence for one part of the theorized causal mechanism. We argue that by establishing an explicit theorized mechanism prior to empirical research and by critically judging our evidence according to an informal Bayesian logic we can remedy some of the problems at hand in much case-study research and increase the inferential leverage in complex within-case evaluation studies.

Keywords
budget support, causal mechanisms, governance, methodology, process tracing, theory-based evaluation

Introduction

In recent years, method scholars have made increased efforts in order to keep up with the evaluative challenges of studying complex policy interventions. Addressing the problem of 'attribution of cause and effect in small n impact evaluations' (White and Phillips, 2012), and calling for the

Corresponding author:
Johannes Schmitt, Institute of Political Sciences, University of Duisburg-Essen; DEval – German Institute for Development Evaluation, Fritz-Schäffer-Str. 26, 53113 Bonn, Germany.
Email: johannes.schmitt@deval.org

'broadening of the range of designs and methods for impact evaluation' (Stern et al., 2012), the debate is driven by the question of how to increase rigor in complex evaluations. Proponents of theory-based evaluations (TBE) have advanced the idea of using a program's theory as an 'explanatory account' of how the program works, thereby focusing analysis on theorized causal mechanisms as a tool to enable stronger causal inferences (Astbury and Leeuw, 2010: 365; Pawson and Tilley, 1997). However, existing theoretical descriptions of TBE do not actually unpack causal processes, tending instead to treat the crucial causal links as 'assumptions' that remain unstudied empirically. Second, while there have been some recent improvements (Lemire et al., 2012), existing TBE approaches still lack a rigorous framework for evaluating the inferential weight of individual pieces of evidence. The result is that existing TBE approaches do not systematically investigate the causal process linking interventions with outcomes (Delahais and Toulemonde, 2012: 291), nor do they enable us to evaluate the strength of the empirical evidence in an evaluation case study.

The articles uses research on the effectiveness of budget support to illustrate some of the challenges faced by existing TBE case studies and the promise of process tracing (PT). With the turn towards the new aid agenda in the early 2000s, development assistance has been increasingly provided in the form of program-based aid.[1] Compared to the former project approach, present-day program aid instruments are described as: complex interventions involving multiple partners, being delivered indirectly through agents, but at the same time they are only a small part of the development portfolio (Stern et al., 2012: 11; see also Betts, 2013: 255). Within the family of program aid, budget support ranges among the most complex development interventions of our times. This ambitious multi-donor aid instrument follows an extensive program logic in which multiple donors provide multiple inputs in order to achieve multiple outcomes (de Kemp et al., 2011). Ever since its introduction, donors have used budget support to pursue multiple objectives. While some focus on the financing objective of budget support to enhance growth and reduce poverty, more and more donors emphasize the governance objective, pointing to the importance of domestic accountability.

Since the introduction of budget support in the early 2000s, donors have increased efforts to evaluate the effectiveness of budget support more systematically. Under the lead of the EU and the OECD-DAC, evaluation experts have adapted a theory-based approach and evaluated budget support against its program logic. While results of numerous case studies suggest that budget support can be effective in promoting macro-economic stability and increasing the share of public expenditure for pro-poor sectors (Caputo et al., 2011; Dijkstra et al., 2012; Rønsholt, 2014; Tavakoli and Smith, 2013), less evidence exists on the instrument's governance effectiveness.[2] We argue that the primary reason for the paucity of evidence-based knowledge stems from methodological weaknesses in the theory-based approach applied in budget support evaluations and show how the use of recent innovations in PT methods can increase the strength of the evidence-based causal inferences we can make when studying complex interventions like budget support. While illustrated using budget support, we believe that the methodological lessons learned are relevant for TBE in general and applicable to evaluations of other types of complex policy interventions.

This article proceeds in two sections. Section 2 illustrates the consequences of the lack of explicit theorization of the causal links in causal processes (mechanisms) in two of the most promising theory-based evaluation approaches (i.e. contribution analysis (CA) and realistic evaluation (RE)) and illustrates the problems using actual case studies from the field of budget support in development aid. The section then explores how these studies could be improved by more explicitly theorizing what it is that links parts of the causal explanation together. Section 3 first exposes the problems related to the logics of inference underlying TBE approaches and existing case studies in the field of budget support. We then present an alternative Bayesian-inspired framework that offers a set of logical tools for evaluating the types and strengths of inferences that pieces of

evidence enable about the underlying causal process one is studying. The article ends with a short conclusion summarizing our methodological recommendations.

Learning from PT when theorizing mechanisms

The theorization of mechanisms in TBE approaches

This section develops the argument that existing TBE approaches lack explicit theorization of the causal mechanisms, and in particular the causal links between parts of mechanisms, therefore providing only a limited base for inferences about actual causal processes as they play out in cases. TBE approaches share the notion of assessing the causal chain from inputs to outcomes and impact against a program logic developed prior to the actual empirical analysis (White, 2009). The a priori theorization of the expected causality chain is described as a strength of TBE and methodological guidance on different methods of reconstructing the program's theory have been provided (Leeuw, 2003).

Mechanisms are described in TBE approaches, notably in RE (Pawson and Tilley, 1997) and more recently in CA (Astbury and Leeuw, 2010; Delahais and Toulemonde, 2012; Leeuw, 2012; Lemire et al., 2012). In practical TBE research however, mechanisms are treated as either a) a series of events or narratives that link the occurrence of X and the occurrence of Y in a particular case (e.g. Astbury and Leeuw, 2010), or b) are theorized in a fashion where the critical causal links in the process are 'grey-boxed' by being treated as 'assumptions' instead of being explicitly theorized as an integral part of the causal story (e.g. Mayne, 2012).

Mechanisms do play a key role in the TBE approach of realistic evaluation (Pawson and Tilley, 1997). Instead of asking 'whether' and 'to what extent' a program contributed to observed outcomes, RE is interested in 'how and why and for whom a program worked'. The realist explanation is that outcome (O) is triggered by a mechanism (M) acting in context (C). Mechanisms explain how a given resource (provided by a program) led to behavioural change in a given context. Prior to empirical research, the most likely CMO configurations are identified in expert interviews (Blamey and Mackenzie, 2007) but the inner parts of the causal mechanism are not theorized *ex ante*. According to realist philosophy, causal processes happen at a different level than the observable program activities and outcomes. Hence, RE does acknowledge that the 'underlying' mechanism is unobservable (see Pawson and Tilley, 1997; Westhorp, 2014) and therefore focuses on identifying the triggers that 'fire the appropriate mechanisms in certain circumstances' (Blamey and Mackenzie, 2007: 446).

Mechanisms are described in more detail in Mayne's Contribution Analysis. Yet, the causal links between parts are treated as assumptions instead of as vital parts of the causal explanation that should be studied empirically. In Lemire et al. (2012) refinement of CA we also see the relegation of these critical causal links to assumptions, meaning that what is actually doing the explanatory heavy-lifting in relation to causal processes is obscured. Leeuw (2012) develops a way of representing 'mechanisms' graphically, but the causal links are also relegated to being undepicted assumptions (Leeuw, 2012: 351). The result is that the ensuing empirical analysis does not study the crucial causal links empirically, widely seen in practical applications of CA (e.g. Biggs et al., 2014; Delahais and Toulemonde, 2012).

In comparison, the causal process between a cause and outcome is made even more explicit in PT due to the detailed theorization of the causal links between causes (X) and outcomes (Y), where the mechanism(s) is disaggregated into a series of interlocking parts composed of entities engaging in activities (see below), whereas in TBE approaches the causal links are relegated to being 'assumptions' or arrows. Yet making all of the links between X and Y explicit exposes the

underlying causal arguments to more scrutiny, resulting in both more logically coherent causal explanations and stronger causal inferences because the analyst is forced to actually study the causal links empirically.

Theorization of mechanisms in existing budget support studies

In practical TBE case studies, mechanisms are often not explicitly theorized. As an example, we deploy the most recent version of a commonly accepted program logic[3] of budget support as presented in the Comprehensive Evaluation Framework (CEF) for evaluating budget support programs (see Figure 1). The CEF builds on earlier work of developing a methodological approach for budget support evaluations (see Booth and Lawson, 2004; EC, 2008; IDD, 2006) and has continuously been substantiated in the course of evaluations over the past decade.[4] This program logic of budget support structures the hypothesized sequence of expected effects across five analytical levels (inputs, direct outputs, induced outputs[5], outcomes and impacts). The CEF allows for various inputs from the government as well as other aid activities and their direct outputs and outlines the external factors and assumptions.

In an earlier description of the program logic, EC 2008 state that '[t]he proposed IL [intervention logic] therefore must incorporate and spell out the anticipated contributions of GBS/SBS to the government strategy and the mechanisms through which GBS/SBS is expected to operate' (EC, 2008: 11). In fact, OECD/DAC (2012) provide a section on the 'main driving forces within the intervention logic of budget support' in which they present two kinds of effects as 'driving forces which generate (most of) the effects at the levels 3, 4 and 5 at the CEF.' (OECD/DAC, 2012: 13). *Flow of funds effects* are expected due to budget support funds provided through the recipient government's own public financial management systems. *Policy and institutional effects* are expected to result from the non-financial inputs of budget support such as policy dialogue between the government and the budget support donors, conditionality, technical assistance and capacity building (TA/CB).[6] Most of these 'driving forces', however, simply represent the effects already presented in the CEF without providing any additional information on the nature of the underlying causal mechanism.

Basically, the program logic depicted here does not develop the causal process whereby individual budget support inputs are expected to work towards the envisaged outputs, outcomes and impacts. The links between the different levels are simply not developed, resulting in a less coherent causal story. And the lack of explicit theorization of causal links has the analytical result that when we do the actual empirical analysis we would not explore how one part leads to the next. We would merely *assume* instead of assess that one part is causally related to the next. Thus, we can only make weaker causal claims than if we explicitly studied the causal process using actual empirical evidence of the causal links between parts of a mechanism.

What PT can offer by taking mechanisms seriously

Recent developments in PT have developed a useful language for theorizing mechanisms, enabling better program theories to be developed. Most importantly, developing how each part of a mechanism is logically linked to each other in terms of entities engaging in activities results in the explicit theorization of causal links instead of relegating them to 'assumptions' (e.g. Lemire et al., 2012). Within the recent literature on PT there is growing recognition that we need to take mechanisms seriously by conceptualizing them as a system that transfers causal forces from cause to outcome.[7] In what can be termed a 'system' understanding of mechanisms the analytical focus is on the generalizable theorized process whereby *causal forces* are *transmitted* through a series of interlocking

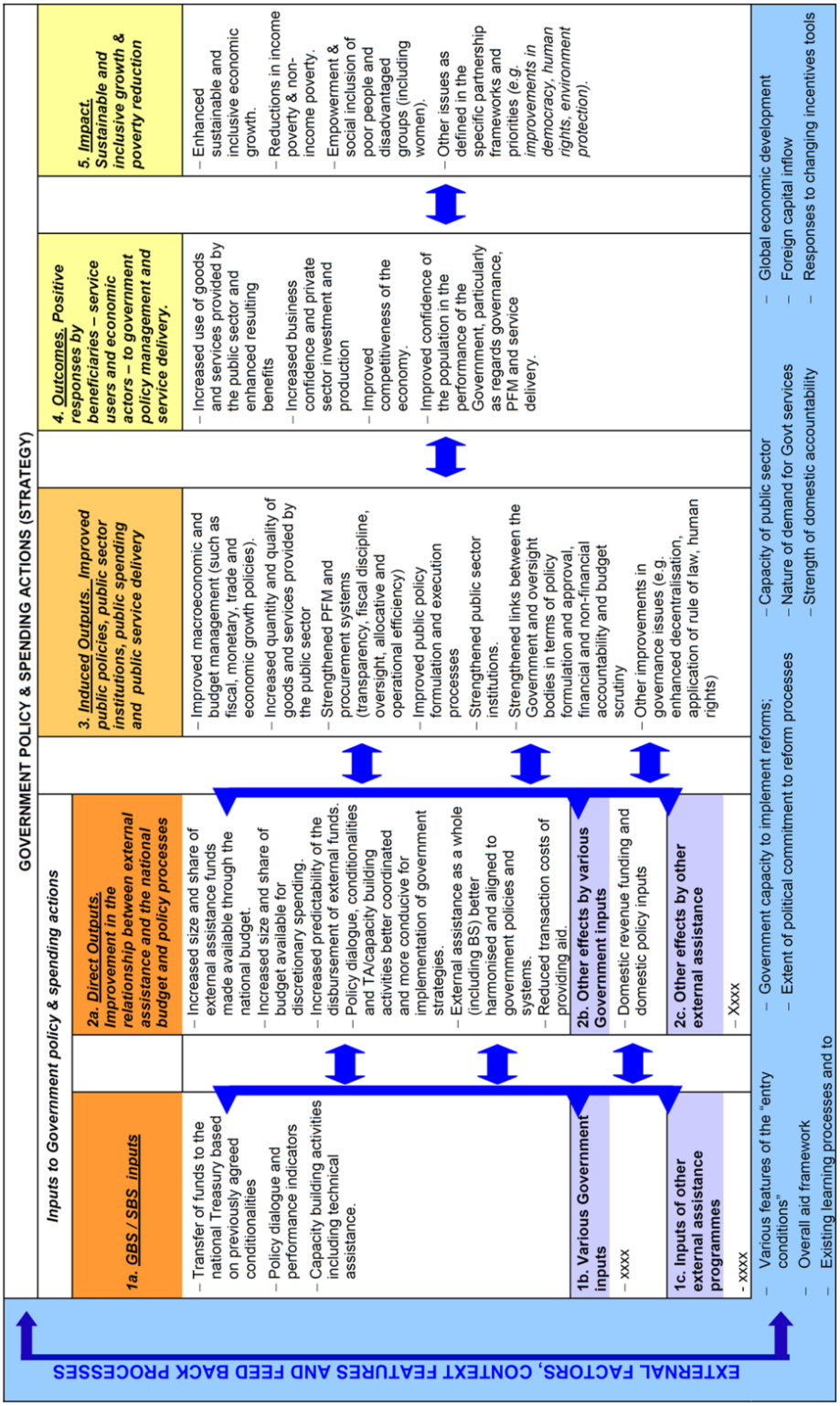

Figure 1. The Comprehensive Evaluation Framework (CEF).
Source: OECD/DAC (2012: 9).

parts of a mechanism to produce an outcome (Bunge, 1997; Glennan, 1996; Mayntz, 2004). Machamer et al. (2000) state that when theorizing a mechanism, we should describe how parts of the mechanism are bound together by 'productive continuity', i.e. there are no logical holes in our explanation of the causal process (Machamer et al., 2000; also Machamer, 2004). They suggest that mechanisms should be disaggregated, viewing them as consisting of a series of parts composed of entities engaging in activities. Entities are what engage in activities, where the activities are the producers of change or what transmits causal forces through a mechanism (Machamer et al., 2000). Parts have no independent existence (i.e. they are not considered as variables) in relation to producing Y; instead they are integral parts of a causal process that *together* produce Y, with each part composed of entities engaging in activities. By explicitly conceptualizing mechanisms and, most importantly, the causal links *between* parts, PT focuses our analytical attention on what it is that links a cause and outcome together in ways that enable us to study the relations empirically.

For mechanisms to operate correctly in a case, the requisite scope (or contextual) conditions need to be present (Astbury and Leeuw, 2010; Blamey and McKenzie, 2007; Falleti and Lynch, 2009; Pawson and Tilley, 1997; Rohlfing, 2014). This context-sensitivity has practical methodological implications for the study of complex interventions such as budget support: we should start by theorizing mechanisms within a specific context unless we have strong cross-case knowledge about scope conditions, detailing only the conditions expected to be necessary for the mechanism to function properly in the particular case. We can then broaden this through the comparative analysis of case studies of budget support interventions, assessing what scope conditions they share and how they differ. In this article we describe a set of assumptions for a simple budget support mechanism to function properly in a particular case (see below).

Another complicating matter is that in complex interventions like budget support, there might be multiple mechanisms triggered by the multifaceted cause, for example direct effects versus more indirect and far-ranging effects, some of which might be quite unintended in an actual case. However, when engaging in PT we typically focus on the link(s) between a cause (X) and a particular outcome instead of multiple different outcomes (Beach and Pedersen, 2013: 14–16). This focus on the mechanism (or mechanisms) linking a particular cause and a single outcome is for analytical simplicity. To keep the analysis manageable, it is best to trace each of the distinct mechanisms separated from each other, unless their effects are so intertwined that we cannot meaningfully separate them. In the example below, we focus our attention more narrowly on the governance effectiveness of budget support inputs and direct outputs at the level of induced outputs. This could then be followed by *additional* PT case studies focusing on other effects of budget support, linking it with other hypothesized outcomes.

Theorizing a causal mechanism for budget support

In the previous section we pointed to methodological shortcomings created by a lack of explicit theorization of causal processes in the program logic of budget support. Taking the same example as used in Figure 1, we hypothesize the theoretical mechanism of how *non-financial inputs of budget support* produce causal effects on the envisaged induced output of *strengthened links between the Government and oversight bodies* and roll out the scope conditions necessary for the mechanism to function. For reasons of analytical clarity, we first need to further specify X and Y. As for X, the program logic states on the level of direct outputs: 'Policy dialogue, conditionalities and TA/capacity building better coordinated and more conducive for implementation of government strategies' (OECD/DAC, 2012: 9). This means that we are dealing with multiple non-financial inputs of budget support. For the Y, it states: 'Strengthened links between the Government and oversight bodies in terms of policy formulation and approval, financial and non-financial

Table 1. A hypothesized causal mechanism for budget support's non-financial inputs.

Part	X: Financial and non-financial inputs of budget support related to oversight bodies are implemented		
	⇓	⇓	⇓
1	Donors and government establish a trustful *policy dialogue*	Donors apply *conditionality* to better integrate oversight bodies into the budget process	Donors provide technical assistance to the staff of oversight bodies to improve capacities
	⇓	⇓	⇓
2	Government provides budget information to facilitate dialogue with donors and send a cooperative signal to secure (additional) financial flows in the preferred form of budget support	Government complies with conditionalities (i.e. improves the formal status of oversight bodies) in order to secure committed budget support funds	Oversight bodies acquire knowledge relevant to their work in budget scrutiny
	⇓		
	The information is shared with different stakeholders (parliament, civil society organizations, media)		
	⇓		
3	The budget process is more transparent (policy dialogue), the formal role of oversight bodies is improved (conditionality) and their capacity to oversee budget processes is increased (technical assistance).→		
4	Better formally integrated and trained oversight bodies increasingly request information on budget issues from the government.→		
5	Driven also by budget support donors, the government provides more and high quality information on budget issues.→		
6	Oversight bodies conduct valuable & solid analysis of budget related issues and provides policy recommendations to the government.→		
7	Government positively reacts to these recommendations and changes its policies.→		
	Outcome (Y): The links between government and oversight bodies are strengthened in terms of budget scrutiny.		

accountability and budget scrutiny' (OECD/DAC, 2012: 9). Oversight bodies here shall include supreme audit institutions, parliament, civil society and the media. For the reason of feasibility, we limit Y to the dimension of budget scrutiny. Focusing on the issue of budget scrutiny the causal mechanism contains multiple parts as displayed in Table 1.

This causal mechanism reflects how different non-financial inputs of budget support are expected to contribute to the envisaged induced output. The mechanism starts with the implementation of the three non-financial inputs of budget support. In contrast to the conception of mechanisms from realist evaluation, in the logic of PT inputs are constituent parts of the mechanism as they entail activities that link each part of the process with each other in a causal sense. They provide the initial impulse for the mechanism to start and are therefore integrated as part 1. An important factor responsible for carrying on the initial impulse in part 2 is the incentive for the government to implement the non-financial inputs. De Kemp et al. (2011: 38) point out that 'by linking policy dialogue and conditionality to funding, the financial contribution also serve to strengthen the effectiveness of non-financial inputs by creating incentives for governance reforms, improvements to policy contents and stronger PFM systems.' With regard to TA/CB, the financial incentive has not been identified to be decisive.[8] The subsequent parts of the mechanisms have been hypothesized

based on a model by Hesselmann (2011). This three-stage procedural model involves actors on the supply side (government entities) and on the demand side of domestic accountability (oversight bodies such as parliament, civil society and the media).

For the mechanism to function, a key assumption and a number of scope conditions have to be fulfilled. In line with existing work on the effectiveness of budget support we assume that national systems are strengthened through the increased flow of funds through the national budget system. Given these 'systemic effects' (Nilsson, 2004) we share the assumption that in the context of budget support, general attention shifts towards the country's own accounting and control systems. In order to make their own budget support contributions more effective, donors use non-financial inputs to directly address capacity constraints and improve domestic accountability systems (Tavakoli and Smith, 2013).

Additional scope conditions on both sides of the aid delivery chain need to be met. On the donors' side, the above mechanism is expected to function best when donors provide well-coordinated and harmonized inputs aligned to the government's strategies and needs. First, the input of policy dialogue can contribute to more transparent budget processes if it is well prepared from the donor side and if clear lines of communications are in place to provide a trustful atmosphere. Second, the ability to establish a credible conditionality framework within the multi-donor setting of budget support is crucial. Only if donors manage to harmonize their individual requirements the conditionality framework can be perceived as a coherent incentive from the government. This aspect has proven to be problematic as the variety of expectations from different donors has at times significantly stretched the conditionality framework to an unfocussed list of indicators and targets (Faust et al., 2012). Third, for TA/CB to be an effective tool to strengthen the capacity of oversight bodies to oversee budget processes, activities need to be well coordinated among donors and fully aligned to the government's needs (Keijzer, 2013; Krisch et al., 2015).

On the recipient side of the budget support relation the quality of national strategies and policies, the government's commitment to actually pursue these policies and the government's political and administrative capacities are decisive for achieving the envisaged governance objective. While these conditions are critical for budget support in general, for the governance objective of increased domestic accountability the supported government needs to be particularly committed to a pro democratic development track irrespective of receiving budget support.

Making stronger causal inferences in evaluation case studies by learning from PT

After the more explicit theorization of causal mechanisms described above, where the causal process is unpacked theoretically, we now turn towards the question of how to generate stronger empirical evidence on the causal process that links causes and outcomes. We illustrate inferential limitations in TBE using examples from budget support evaluations and present innovations from PT (i.e. the informal use of Bayesian logic) as a potential solution.

Existing theory-based approaches to evaluation typically do not describe what is actually doing the analytical heavy-lifting in making evidence-based inferences about causality. In CA, stakeholders are interviewed to find out whether they believe the program worked (Blamey and McKenzie, 2007: 449; Mayne, 2012). However, this would be like trying to take a criminal case to trial only using the testimony of the victim and suspect. Yet real world causal processes (typically) leave a much richer empirical trail, meaning that evidence can be many different types of empirical material. While recent additions to CA like the Relevant Explanation Finder (Lemire et al., 2012) do make progress in describing what types of empirical material can act as evidence, the recent incorporation into PT of Bayesian logic arguably provides us with much stronger logical foundations for

making evidence-based claims using empirical evidence. At the practical level, evaluating the theoretical uniqueness of each piece of evidence introduces a level of control for alternative explanations, thereby avoiding the problems as noted by critics of case studies such as endogeneity and the impact of third causes that often plagues single case studies (Dijkstra and de Kemp, 2015: 87).

The following draws on examples from budget support case studies in the literature. Given the emphasis of this article is on the governance effects of budget support, we focus on the level of induced outputs where most of the governance objectives are spelled out. For the analysis up to this level, two sources of causal inference are discussed in the methodological approach: the use of counterfactuals and the application of process evaluation.

Making inferences using counterfactual thinking in TBE

The use of counterfactuals[9] in evaluation has been used in experimental designs (randomized controlled trials – RCT) enabling the attribution of the observed effect to the controlled intervention. RCTs have been celebrated as the gold standard of rigorous impact evaluations and we acknowledge that they offer powerful, counterfactual-based inferential tools for evaluation. Advocates of the counterfactual approach to causality have had a strong influence on the debate on evaluation designs and 'succeeded in challenging all evaluators to address questions of causal claims and explaining the effectiveness of development interventions' (Stern et al., 2012: 8).

In practice, counterfactual logic is often also applied in non-experimental settings, especially in case studies, where the analyst merely estimates what would have happened if no intervention had taken place. But unless actual manipulations take place, this logic is only a hypothetical 'what if' with no actual empirical evidence backing the inference. In contrast, in PT we are interested in analysing evidence of what *actually* happened and not what could have hypothetically happened. Despite these weaknesses, counterfactual logic has been used in a number of budget support evaluation case studies in a flexible manner (Dijkstra and de Kemp, 2015).

Looking at evaluation case studies shows the flexible use of counterfactuals represented in different ways from case to case and also across aspects covered by the evaluations. In the early set of budget support evaluations conducted by IDD and Associates (2006), counterfactuals have been formulated for each of the sub-inquiries across seven case studies in order to take into account alternative explanations like other aid modalities or different budget support designs. In the final reports, these counterfactuals are addressed in a sub-section for each evaluation question (EQ). However, they are of limited inferential value because they hardly exceed informed guesses on the weight of alternative explanations based on information from interviews (see Batley et al., 2006: 56). In the evaluation of budget support in Zambia, de Kemp et al. (2011) apply a different approach. Instead of comparing the identified outcomes to a situation of other aid modalities, they assess budget support against its own intervention logic and develop counterfactuals at the sector level for selected key issues in order to 'avoid misattribution' (see de Kemp et al., 2011: 45). However, no counterfactuals have been provided ex ante for the analysis of governance effects at the induced output level. Thirdly, the evaluation of budget support in Tanzania does not dwell on the use of counterfactuals in the main report. Despite two counterfactuals considered for the governance-related EQ in the annex (ITAD, 2013: A42), no related information is provided in the main report.

In sum, evaluators have applied counterfactual thinking when conducting budget support evaluations. In order to gain higher credibility, the studies discuss their results against the role of confounding factors like other aid modalities or different policy options. Yet despite their use to increase the quality of the story told, this use of counterfactuals does not provide a viable basis of causal inference given that the inferences have no real empirical evidence backing them, but only logical arguments (Dijkstra and de Kemp, 2015: 88; see also Stern et al., 2012: 8).

Making inferences using 'process evaluation'

Another tool to make inferences in evaluating the effectiveness of budget support inputs uses an inventory of financial and non-financial inputs of budget support as well as a CA of these inputs. This 'process evaluation' (EC, 2008: 18) or 'aid effects evaluation' (OECD/DAC, 2012: 17) is described as 'an analysis of causal relations between inputs and direct and induced outputs. It assesses how [budget support], through its different inputs and mechanisms, has contributed to strengthening government policies, institutions, budget allocation processes, PFM and service delivery' (OECD/DAC, 2012: 17). Unfortunately, however, both EC (2008) and OECD/DAC (2012) remain vague on the methods used to conduct the analysis.

How did evaluations of budget support approach these methodological challenges? In their Mozambique case study, using an earlier version of the evaluation framework, Batley et al. (2006) among other topics, evaluated governance effectiveness by asking the EQ of '[h]ow efficient, effective and sustainable has been the contribution of PGBS to improving government ownership, planning and management capacity, and accountability of the budgetary process?' (Batley et al., 2006: 170). Before conducting empirical research, the evaluation team developed a country-specific causality map in which principle causality chains are displayed.[10] The hypothesized links presented in the report specify additional outcomes without explaining how these outcomes should be reached. The report concludes that 'GBS does appear to be operating through the causality links hypothesised in B4.1' and that budget support could help create conditions for greater accountability by changing the relationship and reporting lines between core government and line ministries (Batley et al., 2006: 57). The reader is however left in the dark on how these results have been produced. Why is the presented empirical material actually evidence that enables causal inferences to be made? The comparatively comprehensive annex on approach and methods elaborates more on instruments (interviews and questionnaires) and reflects problems during the field trips but does not provide information on the judgment criterion and their means of verification used to test the hypothesized causality chains.

The evaluation of budget support in Zambia asks for improvements in governance in EQ 3.5: 'To what extent have there been improvements in governance and democratic accountability, particularly regarding the relative roles of parliament and civil society in relation to the budget?' Being aware of the 'remaining attribution problems', the report carefully concludes that:

> [o]verall the evidence suggests that the PRBS process has helped somewhat to improve policy processes and the overall quality of governance especially with regard to strengthening the supply side of state accountability. Yet, … the PRBS group did not coherently engage in strengthening civil society or the parliament on the demand side of democratic accountability. (de Kemp et al., 2011: 110)

The above conclusion is based on a notable amount of empirical material produced for the levels of inputs, outputs and induced outputs. Qualitative-data processing techniques are made transparent in the report and include the use of coding software for more reliable triangulation of statements by different groups of interviewed stakeholders. The credibility of the governance-related conclusions benefits from the political-economy perspective taken throughout the study. However it remains unclear how empirical material actually has been translated into evidence that substantiates that there are causal relationships in the case. Despite reference to Mayne's CA in the methodology chapter (de Kemp et al., 2011: 44), the report does not give additional information on how and for which EQ this method has been applied. It is not clear to the reader *how* different non-financial inputs of budget support actually contributed to the observed outcomes. The mechanisms remain un-theorized ex ante and not sufficiently assessed empirically. However, at least for the case of the particular question on Parliament and CSO, the hypothetical causal mechanism could

have been developed beforehand based on prior knowledge from theory and empirical studies. Altogether, the evaluation has produced credible results for answering the EQ as quoted above but the conclusions could have included stronger statements on the causal contribution of specific budget support inputs to the observed changes.

Making inferences without correlations or counterfactuals in PT

This section illustrates how recent innovations in PT enable stronger evidence-based causal inferences to be made. In particular, the recent incorporation of the Bayesian logic of inference in PT provides a more rigorous framework for assessing the type and strength of inferences we can make using different forms of empirical evidence (Beach and Pedersen, 2013; Bennett, 2014; Rohlfing, 2014). Used in an informal fashion, where we do not quantify parts of the Bayesian formula, the basic logic shows how we can utilize multiple sources of evidence to update our confidence in the presence/absence of parts of a causal mechanism.

Bayesian logic. The core of Bayesian logic is the intense assessment of what a particular piece of empirical material (evidence) tells us in relation to a theoretical hypothesis in a case. The simplest version of Bayes' theorem is: *posterior = weight of evidence x prior*. This states that our confidence in the validity of a hypothesis is, after collecting evidence (posterior), equal to the probability of the found evidence conditional on the hypothesis being true times the probability that a theory is true based upon our prior knowledge. The goal of empirical tests is to update our confidence in a theoretical hypothesis, although it is always a matter of degree; we never *absolutely* confirm or disconfirm a hypothesis. We increase our confidence when the posterior is greater than the prior, decrease our confidence when the posterior is less than the prior, and learn nothing from our research when the posterior is equal to the prior. Our prior confidence in a theory matters, in that if there are strong existing studies suggesting a theory is valid, only very strong empirical tests will enable further confirmation. In contrast, when we are initially not very confident (low prior), even weak tests can enable positive updating to take place. Note that, in PT, the posterior confidence is only updated in relation to whether the hypothesis holds in the selected case, and any inferences beyond the single case are only made possible by nesting it within a broader comparative design that would for instance enable us to claim that the case was typical.

The analyst starts by making predictions based upon the theory of what pieces of evidence (or empirical fingerprints) we should find in the empirical record. Evidence is understood broadly as any empirical material that has probative value in relation to a given theory. In relation to case studies, we can distinguish between four particular types of evidence. First, *pattern evidence* relates to statistical patterns. If we are testing the overarching claim embedded in the intervention logic of budget support, which is the objective to reduce income and non-income poverty in the recipient country, we should find disproportionately higher shares of the overall government budget to poverty sectors such as health and education *after* the introduction of budget support. A second form of evidence is *sequences*, making predictions about temporal or spatial chronologies of events. When testing the hypothesis at the direct output level that TA/CD activities are better coordinated, relevant evidence might be whether donors assess the technical bottlenecks *before* they assign their TA measures. Third, *trace evidence* refers to material where its mere existence provides proof. If our theory states that oversight bodies have improved knowledge relevant for their work on budget scrutiny due to TA/CD received from budget support donors, we would expect to find budget scrutiny documents produced by the respective oversight body in which methods and techniques acquired in trainings are applied in scrutiny *and* direct reference is made

to the training. Fourthly, *account evidence* relates to material where it is the content that matters. This can be in the form of what participants tell us in interviews, or the content of relevant documents like legislative proposals.

Assessing transparently what causal inferences evidence enables. Where PT departs substantially from TBE approaches like CA is in the explicit assessment of what empirical material can tell us about the causal process being investigated. Here recent developments in PT suggest that we should openly assess: 1) the theoretical certainty of finding the piece of evidence, understood as whether we *have to find it* for the theory to be valid, and 2) the theoretical uniqueness of the piece of evidence, understood as whether there are any plausible alternative explanations for finding it. Certainty captures the disconfirmatory power of a test when we do *not* find the predicted evidence, whereas uniqueness relates to the confirmatory power of found evidence.

Each piece of potential evidence should be assessed in terms of certainty and uniqueness. For example, returning to the overarching claim of budget support to reduce poverty, the analyst should first ask whether we would have to find this evidence of a disproportionate increase of funds allocated to pro-poor sectors to exist in the case (high certainty), or whether budget support might be still valid but that it would not necessarily manifest itself in pro-poor budget statistics (low certainty)? If the prediction was highly certain and we did *not* find it in the case, we would then downgrade our confidence in a part of the mechanism.

Taken together, the levels of theoretical certainty and uniqueness of predicted evidence in relation to a case results in four different types of empirical tests (Beach and Pedersen, 2013; Rohlfing, 2014; Van Evera, 1997). A straw-in-the-wind test combines low certainty and low uniqueness, resulting in little updating. In contrast, a doubly-decisive test is one that combines high uniqueness and high certainty. Not finding the predicted evidence downgrades our confidence, whereas finding it increases our confidence because there are few plausible alternative explanations for the evidence.

There are two types of asymmetric tests, where the confirming power of finding predicted evidence is not the same as the disconfirming power when we do not find the evidence. A hoop test involves making a certain prediction. However, the test is not very unique, in that there are many plausible alternative explanations for finding the evidence. In a hoop test, finding the predicted evidence means little updating takes place, whereas a negative result (i.e. not finding evidence) significantly disconfirms the hypothesis. In contrast, a smoking gun test involves making a theoretically unique prediction but not certain. If found, the test enables strong confirming inferences, whereas not finding the predicted evidence usually does not enable us to conclude anything beyond 'we did not find the smoking gun' (Sobel, 2009: 71). While there is little we can do to alleviate the weakness of smoking gun tests, we can combine multiple hoop tests together to enable confirming inferences to be made (see below).

In a Bayesian framework, research is a cumulative process of updating our beliefs in the validity of theories by employing repeated empirical tests. In particular, multiple (independent) empirical tests have an additive, cumulative effect (Good, 1991). One hoop test by itself might do little to confirm a hypothesis, but when we sum together several hoop tests that do not overlap with each other (i.e. they assess different observable manifestations of a given hypothesis using independent evidence), the final result can be a significant *increase* in our confidence in the hypothesis. In legal reasoning a scale analogy is often used to express the assessment of the inferential weight of multiple tests, with a judgment dependent on whether the preponderance of evidence points in one direction or another.

Developing empirical tests in evaluation case studies

In view of the methodological innovations from PT, we now illustrate the logic by operationalising tests for two parts of the causal mechanism of budget support. We show how to make predictions about what evidence we should find for each part of the mechanism and what types of inference we can draw from the subsequently collected empirical raw material. Equally important, however, evaluators would have to develop an understanding on where specific pieces of evidence can be found and make statements on the probability of observing the evidence ex-ante. Based on the causal mechanism developed above, we state the level of confidence for the hypothesis prior to research for two parts of the causal mechanism and specify the fingerprints we expect it to leave in the case. We then assess what the predicted evidence tells us in relation to the hypothesis, describing what type of empirical test it is.

Tables 2a and 2b show the predicted observable manifestations of evidence for the non-financial budget support input of technical assistance in the form of trainings to representatives of oversight bodies. For two parts of this part of the mechanism, we specify entities and activities and provide a more specific hypothesis for the two parts of the causal mechanism (Table 2a). We state different prior probabilities for the two parts based upon prior knowledge from budget support evaluations and information on the aggregate effectiveness of development projects. While it is very likely to find technical assistance implemented in a case of interest (assuming we choose a case where X – i.e. non-financial inputs – is present), the probability that these technical assistance measures had been successful resulting in improved knowledge relevant for their work on budget scrutiny is moderate.[11]

Given that we can formulate relatively strong disconfirming tests (see Table 2b), two tests should be sufficient to assess part 1 with a reasonable degree of confidence. In a first test, we look for technical assistance described in program documents. Before entering the stage of implementation, technical assistance projects are always planned and described in documents. Hence, this test is arguably quite certain. Uniqueness is however limited to the extent that even if we find written text about the implementation of technical assistance in the document, this does not necessarily confirm our hypothesis because the actual implementation of the planned project could be delayed or cancelled due to technical or political reasons. We therefore include a second test and ask representatives from oversight bodies whether technical assistance has been provided. This doubly-decisive test is strong not only to confirm but also to disconfirm the hypothesis because the source of information (representatives from to be trained oversight bodies) is very close or even part of the target group. The test is certain because the statement needs to be found for h to be true given that we would expect the staff to be regularly informed about trainings. At the same time, the test is unique because finding e (interviewees confirming that technical assistance has been implemented according to the program

Table 2a. Applying the Bayesian logic: hypotheses and probabilities.

	Part	Entity	Activity	Specific hypothesis about causal link	p(h)
	1	Donors	Implement technical assistance	Technical Assistance is being implemented in the area of budget related oversight bodies	Relatively high
	2	Oversight bodies	Receive trainings	Oversight bodies have improved knowledge relevant for their work on budget scrutiny due to technical assistance received	Moderate

Note: p(h): probability for the hypothesis to be true.

Table 2b. Applying the Bayesian logic: test types and predicted empirical evidence.

Part	Means of Verification	Predicted empirical evidence	Description of certainty and uniqueness	Test type
1	Program documents	Technical assistance is described in an official program document and agreed upon by both sides (donor and recipient)	Finding e does not necessarily confirm h given possible delays in project implementation, not finding e disconfirms h	Hoop test
	Interviews with oversight bodies representatives	Interviewees confirm that technical assistance has been implemented according to program document	Finding e confirms h, not finding e disconfirms h	Doubly decisive
2	Interviews with oversight bodies representatives who participated the trainings	Interviewees statements on quality and relevance of training	Finding e does not necessarily increase confidence in h given potential bias of the interviewee to over-rate the quality of the training	Hoop
	Interviews with the trainer	Interviewees statements on quality and relevance of training	Finding e does not necessarily increase confidence in h given potential bias of the interviewee to over-rate the quality of the training	Hoop
	Feedback-forms (from trainings)	Participants rate the training relevant for their work and of high quality	Finding e does not necessarily increase confidence in h given respondents might face a bias (although weaker than in interviews) to over-rate the quality of the training	Hoop
	Budget scrutiny documents produced by the respective oversight body	Methods and techniques acquired in trainings are applied in scrutiny and direct reference is made to the training	Finding e confirms h. Not finding e does not necessarily disconfirm h	Smoking Gun

Note: e: evidence; h: hypothesis that part of a causal mechanism exists.

document) can exclusively be explained by technical assistance actually being implemented. Combining these two tests, we can infer the training was planned (test 1) and actually implemented with the oversight body's representatives (test 2).

Tests for part 2 are less definitive. We would not be able to update our confidence very much based on information from interviews because informants have motivations for not telling us the truth. We would not know what a positive statement on the quality and relevance of the trainings from participants or trainers would stand for. Participants might arguably face the incentive to overstate the value of the workshop or training simply because they benefit financially by receiving per diems from participating in the workshop. Triangulating the participant statements with the

ones made by the trainers might be of limited value as trainers face similar economic incentives and therefore the information provided tends to be biased as well. A way to increase the level of confidence in confirming or disconfirming the hypothesis would be to cross-check evidence with less biased information from sources independent of the interviews. Feedback forms filled in by participants after having received the training might be a good way to triangulate sources given the forms are filled in anonymously. The first three tests provided in Table 2b are relatively weak and would certainly not be enough to substantially increase our confidence in the existence of part 2. More confirmatory power is expected from the fourth test. In this smoking-gun test, finding the piece of evidence (the smoking gun: a budget scrutiny document from the respective oversight body in which the methods/techniques acquired through the training are being applied and direct reference is made to the training) might take more effort than for the other three fingerprints. This test is unique because the occurrence of the predicted piece of evidence cannot be linked to any rival explanation. Due to the unique nature, this test could, in the case of finding the evidence, confirm the hypothesis and prove the existence of part 2 of our causal mechanism. The inferential strength is probably worth the effort of going through the documents.

Conclusion

The core argument of this article is that existing theory-based evaluation methods can be improved by focusing more explicitly on theorizing causal relations, and by improving their inferential underpinnings. Therefore, the arguments made in this article tie in to the debate on how to improve causal inference using theory-based case study designs in evaluations by exploring the methodological value-added brought by recent innovations in the method of Process tracing (PT).

We have illustrated our argument by using examples from the complex development intervention of budget support. Evaluations of budget support have produced an immense body of empirical evidence. However, as we have shown for the particular question on the aid instruments effectiveness on domestic accountability, the inferential power of case studies is limited due to two major reasons. First, there is a lack of explicit theorization of causal mechanisms in the existing intervention logic as applied in several case studies of evaluating complex budget support programs. Notwithstanding being fleshed out as key elements for the empirical analysis, mechanisms are not made explicit ex ante and therefore remain grey-boxed. This under-theorization inadvertently undermines the potential to make causal inferences on the contribution of budget support to the observed outcomes. A second shortcoming is reflected in the lack of a clear logic of inference. We argue that existing theory-based evaluation approaches do not clearly spell out how the evaluator actually knows based on empirical evidence whether an intervention has had an effect. Despite the abundance of empirical information on effectiveness, the inferential basis upon which the causal conclusions are made is often weak. In the exemplified case studies presented above, hypothetical counterfactuals are applied to avoid misattribution and in order to create a better understanding of alternative explanations for particular observed outcomes. Yet, given the non-experimental design of budget support evaluations (no variation, no control group), they cannot be taken as a strong basis for causal inference. A more promising source of inference seems to be what is termed ‘process evaluation’ in the methodological guidelines for budget support evaluations. Unfortunately, however, both EC (2008) and OECD/DAC (2012) remain vague on how to conduct the qualitative analysis and, as the analysis of the two case studies showed, despite high levels of transparency on how the evaluators actually attained and processed empirical evidence, it is not clear how causal conclusions are drawn for the individual evaluation questions.

We have shown that recent innovations in PT methodology can help solve the identified challenges and increase the strength of causal inference made in the demanding settings of complex evaluations. For the specific task of evaluating the governance effectiveness of budget support, we developed a more fine grained causal mechanism for a sub-set of the comprehensive program theory of budget support. Moreover, based on the informal Bayesian logic, we have elaborated on how to increase the conclusiveness of empirical evidence for two parts of the theorized causal mechanism. We argue that by establishing an explicit theorized mechanism prior to empirical research and by critically judging our evidence according to an informal Bayesian logic we can remedy some of the problems at hand and substantially strengthen the inferential lever in complex within-case evaluations.

Notwithstanding the potential benefits from systematically applying PT in the context of complex development evaluations, we also see limitations in using this comprehensive and time-consuming procedure. Given the time and resource constraints in the real world of evaluation, the ex-ante theorization of fine grained causal mechanisms as well as the specification and valuation of their empirical manifestations appears to be a daunting task. Evaluators might rather go for the 'weak but broader' approach instead of overloading the evaluation with troublesome questions of analytical rigor and causal inference. Yet, with more and more voices demanding rigor for qualitative evaluation methods, this 'real world' might change in favor of stronger and more focused approaches to complement the existing broad perspective on complex aid interventions.

An important task for the future will be to break down the comprehensive social science method of PT and make it even more accessible for the hands-on usage in complex evaluations. Further efforts need to be taken to provide a practitioner's guideline on how to apply PT in complex settings and thereby fill the gap in the methodological approach of budget support evaluations.

Notes

1. Program-based aid subsumes aid modalities that support local development programs in a coordinated way. For more information on forms and standards of program-based aid see Pech (2010: 5ff).
2. The term *governance effectiveness* refers to the expected outcomes of budget support in three areas of good governance. Faust et al. (2012: 442f) distinguishes three components: (i) a more effective, transparent, and accountable process of financial planning and implementation, (ii) sector specific institutional reforms and (iii) promoting political processes conducive to democracy. The focus of this paper will be on the third component.
3. We deliberately use the term program logic (interchangeably with the term intervention logic) as opposed to the term program theory. The CEF outlines the anticipated effects on different levels and acknowledges other factors and context. Following the definition by Funnel and Rogers (2011: 31), it cannot be termed a program theory because it does not make explicit the underlying causal mechanisms to explain how the program causes the intended outcomes (see also Rogers et al., 2000: 5).
4. Unfortunately, no specific information is being provided on the process of (re)constructing the program theory.
5. The CEF distinguishes between direct and induced outputs. Direct outputs are 'improvements in the relationships between external assistance and the national budget and policy processes'. Induced outputs comprise 'expected positive changes in the quality of public policies, the strength of public sector institutions, the quality of public spending (increased allocative and operational efficiency), and consequent improvements in public service delivery' (OECD/DAC, 2012: 8).
6. For a complete list of *flow of funds* as well as *institutional effects* see OECD/DAC (2012: 13).
7. This is also termed the 'mechanismic' understanding of mechanisms in the literature. For recent discussions of the mechanisms-as-systems understanding, see Beach and Pedersen (2013); Gerring (2010); Hedström and Ylikoski (2010); Waskan (2011).

8. In a recent evaluation of accompanying measures to budget support, Krisch et al. (2015) find mixed evidence on the incentive effect of budget support on the recipient government's demand for technical assistance.
9. The OECD defines the counterfactual as '[t]he situation or condition, which hypothetically may prevail for individuals, organisations or groups were there no development intervention' (OECD, 2009: 24).
10. The causality chain is presented in the causality map and described in the report (Batley et al., 2006: 51).
11. The probability statements are based on information provided in multiple budget support case studies (Batley et al., 2006; de Kemp et al., 2011; ITAD, 2013) as well as synthesis reports (IDD, 2006; Tavakoli and Smith, 2013). Independent sources on the effectiveness of TA/CD are scarce compared to other types of development assistance (Keijzer, 2013: 2). Taking into account evaluations from implementation agencies, more than three-quarters of development projects are assessed satisfactory or better. In view of the disappointing trends on the macro level, Faust (2010) raises concerns of the effectiveness of development projects being over-rated. Hence, we classify p(h) for part 2 as moderate.

References

Astbury B and Leeuw FL (2010) Unpacking black boxes: Mechanisms and theory building in evaluation. *American Journal of Evaluation* 31: 363–381.

Batley R, Bjørnestad L and Cumbi A (2006) Mozambique Country Report - Joint Evaluation of Budget Support 1994–2004. Available at: http://ec.europa.eu/smart-regulation/evaluation/search/download.do;jsessionid=VQ1QTTbHJgdGv5b5N1QClDD72nh45JYvzhlL5DvGJPBscvRnl44h!1601440011?documentId=2501.

Beach D and Pedersen R (2013) *Process Tracing Methods: Foundations and Guidelines*. Ann Arbor: University of Michigan Press.

Bennett A and Checkel J (2014) *Process Tracing. From Metaphor to Analytical Tool*. Cambridge: Cambridge University Press.

Bennett A (2014) Appendix: Disciplining our conjectures: Systematizing process tracing with Bayesian Analysis. In: Bennett and Checkel (eds) *Process Tracing. From Metaphor to Analytical Tool*. Cambridge: Cambridge University Press.

Betts J (2013) Aid effectiveness and governance reforms: Applying realist principles to a complex synthesis across varied cases. *Evaluation* 19(3): 249–68.

Biggs JS, Farrell L, Lawrence G, et al. (2014) A practical example of Contribution Analysis to a public health intervention. *Evaluation* 20(2): 214–29.

Blamey A and Mackenzie M (2007) Theories of change and realistic evaluation. Peas in a pod or apples and oranges? *Evaluation* 13(4): 439–55.

Booth D and Lawson A (2004) Proposed evaluation framework for general budget support: Framework for country level case studies. Final Report to the OECD-DAC Technical Working Group on Evaluation of Budgetary Aid, ODI.

Bunge M (1997) Mechanism and explanation. *Philosophy of the Social Sciences* 27(4): 410–65.

Caputo E, Lawson A and de Kemp A (2011) Application of new approach to the evaluation of budget support operations: Findings from Mali, Zambia and Tunisia - Synthesis of main results. *Bruxelles: European Commission*. Available at: http://www.oecd.org/countries/mali/49716695.pdf.

De Kemp A, Faust J and Leiderer S (2011) Synthesis report - Between high expectations and reality: An evaluation of budget support in Zambia (2005–2010). Available at: http://www.oecd.org/countries/zambia/49210553.pdf.

Delahais T and Toulemonde J (2012) Applying contribution analysis: Lessons from five years of practice. *Evaluation* 18(3): 281–93.

Dijkstra G and de Kemp A (2015) Challenges in evaluating budget support and how to solve them. *Evaluation* 21(1): 83–98.

Dijkstra G, de Kemp A and Bergkamp D (2012) Budget support: Conditional results - Review of an instrument (2000–2011). The Hague: Policy and Operations Evaluation Department (IOB).

EC (2008) Methodology for Evaluations of Budget Support Operations at Country Level - Issue paper. Available at: http://aei.pitt.edu/47177/.

EC (2012) Budget Support Guidelines Executive Guide - A modern approach to budget support. *European Commission*. Available at: http://ec.europa.eu/europeaid/sites/devco/files/methodology-budget-support-guidelines-executive-guide-201209_en_2.pdf.

Falleti TG and Lynch JF (2009) Context and causal mechanisms in political analysis. *Comparative Political Studies* 42: 1143–66.

Faust J (2010) Wirkungsevaluierung in der Entwicklungszusammenarbeit. *APUZ - Aus Politik und Zeitgeschichte* 10/2010.

Faust J, Leiderer S and Schmitt J (2012) Financing poverty alleviation vs. promoting democracy? Multi-donor budget support in Zambia. *Democratization* 19: 438–64.

Funnell SC and Rogers PJ (2011) *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models*. San Francisco, CA: John Wiley & Sons.

Gerring J (2010) Causal mechanisms: Yes but… *Comparative Political Studies* 43(11): 1499–526.

Glennan SS (1996) Mechanisms and the nature of causation. *Erkenntnis* 44(1): 49–71.

Good LJ (1991) Weight of evidence and the Bayesian likelihood ratio. In: Aitken CG and Stoney DA (eds) *Use of Statistics in Forensic Science*. London: CRC, 85–106.

Hedström P and Ylikoski P (2010) Causal mechanisms in the social sciences. *Annual Review of Sociology* 36: 49–67.

Hesselmann E (2011) The limits of control: The accountability of foundations and partnerships in global health. In: Rushton S and Williams OD (eds) *Partnerships and Foundations in Global Health Governance*. Houndmills: Palgrave Macmillan.

IDD (2006) Evaluation of General Budget Support: Synthesis Report. Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/67831/gbs-synthesis-report.pdf.

ITAD (2013) Joint evaluation of budget support to Tanzania: Lessons learned and recommendations for the future. *Final Report*. Available at: http://www.itad.com/wp-content/uploads/2013/08/Tanzania-BS-Evaluation-Final-Report-Volume-I-180713_AP.pdf.

Keijzer N (2013) Unfinished agenda or overtaken by events? Applying aid- and development-effectiveness principles to capacity development support. *DIE Discussion Paper* 17.

Krisch F, Schmitt J and Doerr U (2015) *Begleitende Maßnahmen der Allgemeinen Budgethilfe in Subsahara Afrika*. Bonn: Deutsches Evaluierungsinstitut der Entwicklungszusammenarbeit (DEval), forthcoming.

Leeuw FL (2003) Reconstructing program theories: Methods available and problems to be solved. *American Journal of Evaluation*. 24: 5–20.

Leeuw FL (2012) Linking theory-based evaluation and contribution analysis: Three problems and a few solutions. *Evaluation* 18(3): 348–63.

Lemire S, Nielsen S and Dybdal D (2012) Making contribution analysis work: A practical framework for handling influencing factors and alternative explanations. *Evaluation* 18(3): 294–309.

Machamer P (2004) Activities and causation: The metaphysics and epistemology of mechanisms. *International Stdies in the Philosophy of Science* 18(1): 27–39.

Machamer P, Darden L and Craver CF (2000) Thinking about mechanisms. *Philosophy of Science* 67(1): 1–25.

Mayne J (2012) Contribution analysis: Coming of age? *Evaluation* 18: 270–80.

Mayntz R (2004) Mechanisms in the analysis of social macro-phenomena. *Philosophy of the Social Sciences* 34(2): 237–59.

Nilsson M (2004) Effects of budget support: A discussion of early evidence. *UTV Working Paper* Stockholm: SIDA.

OECD (2009) DAC glossary of key terms and concepts. Available at: http://www.oecd-ilibrary.org/development/development-co-operation-report-2009_dcr-2009-en.

OECD/DAC (2012) Evaluating budget support: Methodological approach. Available at: http://www.oecd.org/dac/evaluation/dcdndep/Methodological%20approach%20BS%20evaluations%20Sept%202012%20_with%20cover%20Thi.pdf.

Pawson R and Tilley N (1997) *Realistic Evaluation*. London: SAGE.

Pech B (2010) Programmorientierte Gemeinschaftsfinanzierung: Implikationen für Post-Konflikt-Situationen (Literaturstudie). *Project Working Paper*. Duisburg: Institut für Entwicklung und Frieden.

Rogers PJ, Petrosino A, Huebner TA, et al. (2000) Program theory evaluation: Practice, promise, and problems. *New Directions for Evaluation* 87: 5−13.

Rohlfing I (2014) Comparative hypothesis testing via process tracing. *Sociological Methods and Research* 43(4): 606−42.

Rønsholt FE (2014) *Review of Budget Support Evaluations*. Kopenhagen: DANIDA.

Sober E (2009) Absence of evidence and evidence of absence: evidential transitivity in connection with fossils, fishing, fine-tuning, and firing squads. *Philosophical Studies* 143(1): 63–90.

Stern E, Stame N, Mayne J, et al. (2012) Broadening the range of designs and methods for impact evaluations. Report of a Study Commissioned by the Department for International Development (DFID).

Tavakoli H and Smith G (2013) Back under the microscope: Insights from evidence on budget support. *Development Policy Review* 31: 59−74.

Van Evera S (1997) *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.

Waskan J (2011) Mechanistic explanation at the limit. *Synthese* 183: 389−408.

Westhorp G (2014) *Realist Impact Evaluation – An Introduction. ODI Methods Lab*. London: Overseas Development Institute (ODI).

White H (2009) Theory-based impact evaluation: principles and practice. *Journal of Development Effectiveness* 1: 271−84.

White H and Phillips D (2012) Addressing attribution of cause and effect in small n impact evaluations: Towards an integrated framework. Working Paper 15. New Delhi: International Initiative for Impact Evaluation.

Johannes Schmitt is a PhD candidate at the University of Duisburg-Essen, Germany and is doing evaluation work in the field of budget support.

Derek Beach is an associate professor at the University of Aarhus, Denmark. He has co-written a book on Process-tracing methods (University of Michigan Press).