Integrating theory-based evaluation and process tracing in the evaluation of civil society gender budget initiatives

Reference:

Bamanyaki Patricia, Holvoet Nathalie.- Integrating theory-based evaluation and process tracing in the evaluation of civil society gender budget initiatives

Evaluation : the international journal of theory, research and practice - ISSN 1461-7153 - 22:1(2016), p. 72-90

Full text (Publishers DOI): http://dx.doi.org/doi:10.1177/1356389015623657

Integrating theory-based evaluation and process tracing in the evaluation of civil society gender budget initiatives.

Reference:

Bamanyaki, Patricia and Holvoet, Nathalie (2016) “Integrating theory-based evaluation and process tracing in the evaluation of civil society gender budget initiatives”, *Evaluation* 22 (1): 72-90

DOI: 10.1177/1356389015623657 evi.sagepub.com

Introduction

The World Health Organisation (WHO) estimates that 99 per cent of maternal deaths worldwide occur in developing countries, particularly Sub Saharan Africa and South East Asia, and most deaths are largely preventable (WHO, 2014). Over the decades, especially following the incorporation of MDG 5 (to improve maternal health) at the United Nations Summit in 2000, numerous strategies and interventions have been implemented by governments and independent organisations to reduce maternal mortality and achieve universal access to reproductive health (WHO, 2014). Gender-responsive budgeting (GRB) has emerged as an effective tool for governments, premised on the argument that commitments towards gender equality and the realisation of women's rights need to be backed by appropriate funding in order to be implemented successfully (United Nations Population Fund and United Nations Development Fund for Women, 2006).

GRB does not imply the production of separate budgets for women. Rather, it is a systematic assessment of any form of public expenditure or revenue raising measure for its impacts on women and girls as compared to men and boys (Elson, 2002), followed by

informed actions to change policies and budgets so as to promote gender equality (Sharp, 2007). GRB initiatives purpose to ensure that the specific needs and interests of women, men, boys and girls belonging to different social groups are adequately considered in decision-making regarding public expenditure and revenue generation (Sodani and Sharma, 2008).

Regardless of the common objective, GRB initiatives vary across countries in terms of political location (inside- or outside-government), focus (national, sub-national or local levels), entry points to the budget (planning, appraisal, audit or evaluation stages), scope (the whole budget, selected sectors or budget items), tools used for analysis and participants involved, among others (Elson, 2002; Budlender and Hewitt, 2006; Sharp and Elson, 2012). The diverse nature of GRB initiatives classifies them as complex interventions with multiple actors, 'multi-faceted processes', varying outputs and effects and no single means to evaluate success (Sharp and Elson, 2012: 1).

Carlitz (2013) notes that while more than 100 countries have implemented GRB initiatives, limited evidence exists of their effectiveness and impact. A synthesis review by McGee and Gaventa (2010) highlights that most of the existing literature is policy- or practitioner-

oriented, largely descriptive and very few comparative studies discuss and explain the degree of effective implementation. In the last decade, available evaluations of GRB initiatives conducted across countries have mostly comprised studies commissioned by international organisations such as the United Nations Development Fund for Women (UN Women, 2010) to evaluate programme implementation of supported initiatives as well as progress towards achieving GRB programming outputs and outcomes. According to Combaz (2013: 3), the limited evidence is attributed to complexities in assessing and interpreting impact, notably variations in the definition and scope of GRB, diversity and unevenness of GRB implementation, "causalities are complex, multiple and difficult to establish" and the effects of GRB may take long to emerge.

This article attempts to address these issues by arguing for the integration of theory-based evaluation with process tracing methods to empirically evaluate the effectiveness and impact of complex GRB initiatives. We utilise a case study of the Forum for Women in Democracy (FOWODE) GRB initiative in Kabale District, rural Uganda and demonstrate the application of this integrated approach to evaluate the effects of the GRB initiative on gender-responsive maternal health service delivery.

The article proceeds with a short overview of theory-based evaluation, process tracing and the integrated approach in Section two. Section three delves into the application of the integrated approach, starting with an elaboration of the programme theory of how local-level civil society GRB initiatives are expected to influence gender-responsive maternal health service delivery. On the basis of the postulated links in the theory of change, a case-specific causal mechanism is hypothesised, along with empirical predictions of observable implications of the mechanism in the case. Next, the empirical evidence found in the case is evaluated using Bayesian logic to make inferences about the presence of specific parts of the causal mechanism. Section four concludes the article.

An overview: theory-based evaluation, process tracing and the integrated approach

Theory-based evaluation (TBE) and process tracing methodology (PTM) are theory-centred approaches that seek to explain how and why a programme realises (or fails to achieve) results (Birckmayer and Weiss, 2000). This section briefly explains TBE and PTM and builds an argument for integrating the two approaches to strengthen the causal claims made about the effects of GRB initiatives on maternal health service delivery.

TBE

TBE, also known as programme theory or theory of change, refers to 'an explicit theory or model of how the programme causes the intended or observed outcomes and an evaluation that is at least partly guided by this model." (Rogers et al., 2000: 5). TBE approaches have grown in popularity over the years, inspired by contributions from numerous authors such as Suchman (1967), Weiss (1972, 1995), Chen (1990) and Rogers et al. (2000). TBE is particularly useful in the evaluation of complex interventions where outputs and outcomes cannot be easily identified or measured directly (Hickey et al., 2015).

Diverse authors have applied TBE differently depending on the evaluation purpose. Weiss (1997) distinguishes two types of TBE according to the use of programme theory. The first type, implementation theory, is concerned with how a programme is carried out and utilises programme theory to evaluate implementation failure or success (Weiss, 1997). The second type, programmatic theory, is concerned with mechanisms – 'the response that activities generate' – between the delivery of programme service and the occurrence of anticipated outcomes (Weiss, 1997: 46). Weiss's distinction is analogous to Rogers et al. (2000), who differentiate between programme evaluations that are intended for causal attribution and those that use the programme theory as a guide to assess programme implementation.

TBEs focused on causal attribution have predominantly combined programme theory with experimental or quasi-experimental designs to identify and measure intermediate steps of programme implementation and proximal outcomes so as to make judgements about programme success (Rogers et al., 2000; Stern et al., 2012). With the increasing complexity of real-world interventions, counterfactual designs have been found to be inappropriate (Stern et al., 2012; Befani and Mayne, 2014) and recent years have seen the rise of non-counterfactual designs being applied in 'small-n' studies to evaluate the effects of complex interventions (White and Phillips, 2012). White and Phillips (2012: 7) identify four types of evaluation approaches that draw on implicit theories 'to establish beyond reasonable doubt how an outcome or set of outcomes occurred', namely realist evaluation (RE) (Pawson and Tilley, 1997), contribution analysis (CA) (Mayne, 2012), general elimination method (GEM) – modus operandi – (Scriven, 2008) and process tracing (George and Bennet, 2005; Beach and Pedersen, 2013).

RE conceives programmes as 'theories incarnate' and maintains that programmes initiate mechanisms which are 'fired up' in given contexts to generate particular outcome patterns (Pawson and Tilley, 1997). RE tests different context-mechanism-outcome pattern

configurations to establish a mid-range theory that explains 'what works for whom, in what circumstances and in what respects and how' (Pawson and Tilley, 1997: 2). CA postulates that a single intervention 'is unlikely to be the sole cause of a subsequent change' and tests the programme theory, its underlying assumptions, as well as other influencing factors to establish whether an intervention is an important contributory cause of the observed change (Befani and Mayne, 2014: 20). GEM generates a list of possible causes of an outcome of interest along with a list of modus operandi (footprints) for each listed possible cause (Scriven, 2008). The facts of the case are then examined to systematically eliminate the possible causes that do not fit with the modus operandi present in the case, leaving a dominant cause or set of causes that explain how the outcome resulted (Scriven, 2008).

RE, CA and GEM are attempts to provide rigorous accounts of how and why an intervention contributed to producing the observed effects. The three approaches, however, do not explicitly unpack the 'causal links in causal processes' and either treat underlying mechanisms as unobservable (RE); or as 'assumptions instead of as vital parts of the causal mechanism that should be studied empirically' (CA) (Schmitt and Beach, 2015: 430-431); or are silent about causal mechanisms (GEM). PTM (discussed in the next section) is increasingly gaining popularity as an alternative approach that provides a solution to this

problem, thereby enabling stronger causal inferences to be made about complex interventions (Schmitt and Beach, 2015; Befani and Mayne, 2014; Stern et al., 2012). PTM, like CA, acknowledges that a particular outcome may be produced by different mechanisms. PTM, however, does not 'test the relative explanatory power of competing mechanisms against each other', but rather controls for plausible alternative explanations for the outcome using Bayesian logic (Beach and Pedersen, 2013: 90) as will be discussed later.

PTM

PTM is a tool for within-case qualitative data analysis and refers to the 'systematic examination of diagnostic evidence selected and analysed in light of research questions and hypotheses posed by the investigator' (Collier, 2011). PTM hypothesises a causal mechanism that is believed to explain how a cause or set of causes contribute to producing an observed or intended outcome and sets out to confirm the existence of the mechanism by 'observing whether case-specific implications of its existence are present in the case' (Beach and Pedersen, 2013: 15).

A causal mechanism can be viewed as a series of interlocking parts comprising entities engaging in activities to transmit causal forces from an initial cause X to the final outcome Y (Beach and Pedersen, 2013). Activities in a mechanism are thought to be the producers of change, while entities are the factors that engage in activities (Machamer et al., 2000). Each part of the mechanism is conceptualised as being individually insufficient but necessary, as it functions together with other parts of the mechanism to produce the outcome (Beach and Pedersen, 2013). Consequently, a causal mechanism can only be confirmed to have been present in a case (with some degree of confidence) if there is strong evidence that all parts of the mechanism were present and functioned as predicted (Beach and Pedersen, 2013).

Beach and Pedersen (2013) prescribe a three-step procedure for testing theories using PTM so as to make strong causal inferences about observed effects. Step one involves transforming causal theories into a clear hypothesised mechanism of how a particular outcome is produced (Beach and Pedersen, 2013). This entails defining the cause (or set of causes), X, the outcome, Y, the theoretical process linking X to Y and the necessary scope conditions for the mechanism to operate correctly (Beach and Pedersen, 2013). In step two, the theorised mechanism is made operational by predicting pieces of evidence (observable

manifestations) that we should expect to find for each part of the mechanism if the mechanism is present in the case (Beach and Pedersen, 2013). Van Evera (1997) introduced the terminology of 'uniqueness' and 'certainty' to describe two types of empirical predictions of evidence. Unique predictions refer to 'empirical predictions which do not overlap with those of other theories' (Beach and Pedersen, 2013: 101). Certain predictions are unequivocal and we must observe the evidence or the theory fails the empirical test (Beach and Pedersen, 2013).

Lastly in step three, relevant evidence is collected and tested to evaluate the presence of each part of the mechanism and the mechanism as a whole having functioned to contribute to producing the observed effects (Beach and Pedersen, 2013). Four types of evidence are distinguished as relevant for case studies, namely: (a) statistical patterns in the evidence, (b) sequence – temporal or spatial chronology of events, (c) trace – evidence whose mere existence proves that a part of the theorised mechanism exists, and (d) account – the content of the empirical material (Beach and Pedersen, 2013: 99-100).

Uniqueness tests – smoking gun tests – confirm the functioning of a hypothesised part of a mechanism based on the 'signature' evidence it leaves behind that is deemed to be 'unique

to that mechanism and practically impossible to have been left by other mechanisms' (Befani and Mayne, 2014: 24). Certainty tests – hoop tests – disconfirm hypotheses, as they enable the ruling out of some mechanisms if the evidence is not found (Befani and Mayne, 2014). When multiple independent hoop tests are combined to test a hypothesis, the result is 'an additive effect that increases our confidence in the validity of (h) [the hypothesis] given that the probability of non-valid hypotheses surviving multiple independent hoop tests fails after each successive hoop' (Beach and Pedersen, 2013: 105). Two other empirical test types are the 'straw-in-the-wind', which neither confirms nor disconfirms a hypothesis; and 'doubly-decisive' which simultaneously confirms the hypothesis and rejects all its other alternative hypotheses (Befani and Mayne, 2014). Table 1 below summarises the attributes of the different tests.

[INSERT TABLE 1 HERE]

PTM employs Bayesian logic to evaluate whether 'finding specific evidence confirms/disconfirms a hypothesis that a part of the mechanism exists relative to the prior expected probability of finding this evidence' (Beach and Pedersen, 2013: 83). In other words, we seek to update our confidence in the likely truth of a hypothesis conditional on finding new evidence, referred to as the posterior probability (p(h|e)) (Bennet, 2014; Beach

and Pedersen, 2013). Three pieces of information are required to estimate the posterior probability namely: (1) 'our initial confidence that a theory is true even before looking at new evidence', referred to as the prior probability (p(h)); (2) the likelihood that 'if a theory is true in a case, we will find a particular kind of evidence in that case' (p(e|h)); and (3) the likelihood that we would find the same evidence in the case if the theory of interest was false (p(e|~h) (Bennet, 2014: 278). The application of Bayesian logic to confirm or disconfirm parts of the hypothesised mechanism is presented later in this article.

The integrated approach: TBE-PTM method

The integrated approach combines TBE with PTM in a single case study to empirically evaluate the research problem and make stronger within-case causal inferences. We argue that this approach helps to address the challenges put forward by Combaz (2013) (see introductory section of this article) that account for the limited evidence of the effectiveness and impact of GRB initiatives.

Given the diverse nature of GRB initiatives, TBE enables us to develop a programme theory that is applicable to the specific intervention under evaluation. TBE utilises

information from various sources in the development of the programme theory – notably interviews with programme architects, managers, practitioners; reviews of programme documents and other relevant policy and academic literature; site visits to observe and understand the context of programme implementation – providing us with a comprehensive understanding of the causal chain of events from the GRB initiative inputs and activities to the intended outcome (influencing gender-responsive maternal health service delivery), along with relevant assumptions underlying the theory (White, 2000).

TBE approaches, however, do not make explicit the causal links between events in the causal chain from the intervention's inputs to the intended outcome, leaving us in the dark about the actual causal forces responsible for producing the outcome (Delahais and Toulemonde, 2012; Befani and Mayne, 2014; Schmitt and Beach, 2015). A review of evaluations conducted to assess the relevance and effectiveness of UN Women's work in GRB between 2008 and 2011 reveals the use of a descriptive TBE approach, with the focus placed on evaluating stated and implicit assumptions that affect programme development (UN Women, 2010). Schmitt and Beach (2015: 430) highlight two weaknesses of such TBE approaches as being the failure to 'unpack causal processes, tending to treat the crucial causal links as "assumptions" that remain unstudied empirically' and (2) the lack of

‘a rigorous framework for evaluating the inferential weight of individual pieces of evidence.’

PTM makes the assumed links in the programme theory more explicit by theorising a case-specific plausible causal mechanism that develops how each part is logically linked to subsequent parts in terms of entities engaging in activities (Schmitt and Beach, 2015: 432). As suggested by Beach and Pedersen (2013), PTM enables us to capture the transmission of causal forces whereby X contributes to producing Y, forcing us to investigate not only inputs, outputs and outcomes in the causal chain, but also the theoretical process linking X to Y. PTM further provides clear guidance on the type of evidence to be collected and relevant criteria with which to judge the strength of the evidence found in the case (Befani and Mayne, 2014). Additionally, the empirical evaluation of evidence using Bayesian logic strengthens our ability to confirm or disconfirm (with a reasonable degree of confidence) the presence of the causal mechanism linking the GRB intervention to the observed changes in maternal health service delivery (Beach and Pedersen, 2013).

We conclude this section by underscoring the usefulness of the TBE-PTM approach in evaluating the effectiveness and impact of complex interventions such as GRB initiatives

and other gender-focused interventions. Gender-focused interventions involve diverse stakeholders engaged in multiple actions and processes and operating in different contexts to bring about social, political and cultural transformation that promotes gender equality. TBE-PTM is a context-specific approach that permits the articulation of a shared theory of how the activities of the various stakeholders contribute to producing the desired changes, along with the mechanism through which causal forces are transmitted to produce these changes. The strength of this approach is that it enables 'articulation and questioning of myths' about gender mainstreaming by pointing out how transformation systematically occurs (Van Eerdewijk and Brouwers, 2014: 5). This approach, therefore, can make explicit the challenges that have hindered gender mainstreaming interventions from achieving success and inform future programming for enhanced results. With specific regard to maternal health, TBE-PTM facilitates an in-depth understanding of how GRB activities such as the inclusion of women in public decision-making forums transforms gender relations and influences women's capacity to change public policies and budget allocations in favour of their health needs (Hofbauer and Garza, 2009). The application of this integrated approach is illustrated in the next section.

Applying the integrated approach: TBE and PTM

This section applies TBE and PTM to evaluate the effects of the FOWODE GRB initiative on maternal health service delivery in Kabale District. The integrated approach proceeds as follows. First, TBE is applied to explicate an implicit programme theory of how local-level civil society-led GRB initiatives are expected to influence improvements in gender-responsive maternal health service delivery, along with the underlying assumptions for the theory to work. Second, the three-step process tracing procedure (PTM) is applied, beginning with: (a) theorisation of a plausible causal mechanism based on the causal links postulated in the programme theory; (b) operationalisation of the mechanism by making case-specific predictions of observable manifestations we would expect to find in the case if the mechanism is present along with relevant empirical tests based on Bayesian logic; and (c) evaluation of the found evidence in the case for each part of the mechanism using Bayesian logic to make inferences about whether we can update our confidence about the presence of individual parts of the mechanism functioning to produce the outcome (Beach and Pedersen, 2013: 91). For purposes of demonstrating the TBE-PTM approach in a little

more detail, we limit the three-step process tracing procedure to two parts of the causal mechanism at district level.

Qualitative data for this study was collected by the first and main author in Kabale District from September 2014 to January 2015. Information drew from four sources, including semi structured interviews with 30 individuals comprising technocrats, district councillors, health workers, media journalists, FOWODE staff and staff from non-government organisations (NGOs) engaged in the maternal health sector of Kabale district. Two focus group discussions were held with grassroots community group members and women of reproductive age in Kamwezi Sub County (where FOWODE implemented its grassroots-level GRB initiative). Other information sources were reviews of FOWODE programme documents and reports, national and district official documents and reports, media recordings and articles, and observation of processes and physical outputs (where possible) to corroborate interview statements. NVIVO 10 software was used for thematic content analysis.

Explicating the programme theory

The programme theory presented in this section is a programmatic theory (Weiss, 1997), as it elaborates the causal-chain responses between the activities of the civil society-led GRB initiative and the occurrence of the anticipated changes to maternal health service delivery at local government level. The programme theory utilises existing literature on civil society-led budget initiatives (Goetz and Jenkins, 1999; McGee and Gaventa, 2010; Robinson, 2008) as well as opinions from various stakeholders, including functional experts, civil society GRB practitioners at national and local level in Uganda to develop a coherent theory. The programme theory depicts a civil society NGO with GRB interventions at district level targeting policy makers (councillors), policy implementers (technocrats) and citizens at the grassroots (parish) level.

The programme theory starts with the broad assumption that lack of awareness among the legislature, executives and health service providers of gender-specific maternal health issues affecting communities and GRB, coupled with the lack of citizen participation

(especially women) in planning and budget processes affects the development and implementation of responsive maternal health policies, budgets and services (Jahan, 1996).

In response to these gaps, it is theorised that the civil society NGO implements a two-pronged strategy that targets district-level actors and grassroots community members (see Figure 1 below). Starting with the district-level intervention, the civil society GRB initiative trains district-level technocrats and female councillors on GRB, provides technical support to technocrats on GRB, and conducts and disseminates findings of independent gender analyses of district health budgets to technocrats (box 1a). This is expected to increase their awareness of the prevailing gender inequalities in the health sector, including maternal health issues, and enhance their capacities to promote and/or ensure gender equity in health policies, budgets and service delivery (box 2a). With increased awareness and enhanced capacities in GRB, it is expected that technocrats and councillors will prioritise and integrate a gender perspective in health policies and budgets and increase internal monitoring of health service provision for gender-responsiveness and accountability (box 3a). The latter two actions are expected to result in improved and gender-responsive maternal health service delivery (box 4).

[INSERT FIGURE 1 HERE]

At grassroots level, it is theorised that the civil society NGO mobilises and sensitises citizen groups (mainly women) on gender, advocacy, health rights, local government planning and budget processes, and simplifies and disseminates health budget information and service standards for popular consumption (box 1b). These interventions are expected to increase awareness and understanding among grassroots-level citizens of gender equality, health rights, local government planning and budget processes, budgets and health service standards (box 2b). With increased awareness and understanding, it is expected that grassroots citizens (especially women) will actively engage in local level planning and budget processes, tracking of public health budgets and monitoring of maternal health service delivery for responsiveness and gender accountability (box 3b).

The interface between politicians, technocrats, health service providers and grassroots-level citizens in public decision-making arenas and through the media (box 3b) is expected to affect prioritisation and integration of maternal health needs in policies and budgets and increased internal monitoring of health service provision (box 3a), leading to improved and gender-responsive maternal health service delivery (box 4) respectively (Papp et al., 2013).

The interface between health providers and citizens during citizen monitoring of health services at health facilities is expected to provide health workers with a better understanding of citizen needs, leading to improved and gender-responsive maternal health service delivery illustrated by the arrow between boxes 3b and 4.

The programme theory above additionally presumes two reverse causal effects at district level and at grassroots level respectively. It is expected that the practice of making gender-sensitive health policies, plans and budgets and the monitoring of health service provision by technocrats and politicians (box 3a) will further enhance their awareness and capacities to promote gender equity in the health sector (box 2a). Similarly, grassroots citizen participation in planning and budget processes, tracking health budgets and monitoring health service provision for accountability (box 3b) is expected to further enhance their understanding of government budgets, health rights and entitlements (box 2b), with the potential effect of increasing their demand for skilled maternal health care.

According to existing literature on budget advocacy initiatives (Goetz and Jenkins, 1999; McGee and Gaventa, 2010) the following supportive conditions (underlying assumptions) are advanced for successful implementation of GRB: (1) openness, sensitivity, political will

and commitment of politicians and technocrats towards gender equity and health sector issues; (2) willingness, capability and commitment of citizens (especially women) to actively participate in public processes; (3) enabling legal, political and cultural environment for citizen participation (especially women) in public processes; (4) institutional mechanisms to enforce compliance among health service providers; and (5) a well-functioning health system with the capacity to respond to citizen demands. The underlying assumptions provide an insight of the scope conditions to look out for in the specific case.

Theorising the causal mechanism

As the first step of the theory-testing process tracing procedure, this section makes more explicit the postulated links between the boxes in the programme theory (see Figure 1) by theorising a three-part case-specific causal mechanism. The causal mechanism is intended to capture the transmission of causal forces from the GRB intervention activities to the intended outcome of improved and gender-responsive maternal health service delivery. Part 1 of the mechanism theorises the causal link between the activities of the FOWODE GRB initiative and the proximate outcomes at district level (boxes 1a and 2a) and at grassroots

level (boxes 1b and 2b) in Figure 1 respectively. Part 2 of the mechanism theorises the causal link between the proximate outcomes and intermediate outcomes at district level (boxes 2a and 3a), at grassroots level (boxes 2b and 3b), as well as the horizontal causal link between intermediate outcomes at grassroots level and at district level (boxes 3b and 3a) in Figure 1 respectively. Lastly, Part 3 theorises the causal link between the intermediate outcomes at district and grassroots level (boxes 3a and 3b) and the distal outcome of improved and gender-responsive maternal health service delivery (box 4) in Figure 1.

In line with the programme theory, we define the initial set of causes (X) as FOWODE GRB advocacy and capacity building interventions at district level and grassroots level aimed at fostering gender accountability in the health sector. As outlined in the programme theory above, these interventions involve training technocrats and female councillors on GRB; providing technical support to technocrats on GRB; and dissemination of independent gender analyses of district health budgets to stakeholders at district level. At grassroots level, the interventions involve mobilisation and sensitisation of women-dominated citizen groups on gender advocacy, health rights and local government planning and budget processes; and simplification of budget information and health service standards

for popular consumption. The outcome (Y) denotes the influence of FOWODE's GRB initiative on improved and gender-responsive maternal health service delivery. The causally relevant dimensions of improved and gender-responsive maternal health service delivery relate to local government responses to address gender-related challenges associated with accessibility (service location), availability (skilled birth attendants, medical supplies, round-the-clock efficient services), affordability (costs) and acceptability (services and staff attitudes) (Jacobs et al., 2012).

The theoretical process linking civil society-led GRB initiatives to gender-responsive maternal health service delivery draws from existing literature on the Social Accountability theory. According to the theory, policy makers, politicians, service providers and citizens are linked in relationships of power and accountability (World Bank, 2005). Citizens exercise *voice* over politicians and policy makers through formal mechanisms (political parties, elections) and through informal mechanisms (advocacy campaigns, public demonstrations and protests) (World Bank, 2005). Politicians and policy makers set directions and provide incentives for service providers to operate in a *compact* relationship. Where specified in the compact, politicians and policy makers also reward or penalise service providers depending on their services and output (World Bank, 2005). Organisation

providers (ministries, departments or agencies in the various sectors) set internal policies and regulations specific to their organisation and manage the operations of frontline providers who provide services to citizens (World Bank, 2005). Lastly citizens, in their role as clients, exercise *client power* over frontline providers through their interactions and monitoring of frontline provider actions (World Bank, 2003).

The necessary scope conditions refer to the context under which the causal mechanism is theorised to function. When studying complex interventions like GRB initiatives, Schmitt and Beach (2015: 434) propose that mechanisms should be theorised within a specific context unless there is 'strong cross-case knowledge about scope conditions'. We theorise the scope conditions necessary for the mechanism as a whole to operate in the case as: (1) an enabling legal, political and cultural environment for citizen participation (especially women) in public decision-making processes; and (2) institutional mechanisms exist to enforce compliance among health service providers.

Having defined X, Y and the theoretical process linking X to Y, we hypothesise that FOWODE, a civil society GRB initiative with district- and grassroots-level advocacy and capacity building interventions influenced gender-responsive maternal health care in

Kabale District through a voice and accountability causal mechanism consisting of three parts, each of which is individually insufficient but a necessary part of the whole mechanism. The initial conditions preceding the causal mechanism are that the legislature, executives and health service providers in Kabale District lack awareness of GRB and gender-specific maternal health issues affecting communities and there is lack of citizen participation (especially women) in public decision-making processes. The causal mechanism is illustrated in Figure 2 below.

[INSERT FIGURE 2 HERE]

The above mechanism forms our whole hypothesis (H). Plausible alternative explanations to this hypothesis that could have influenced improved and gender-responsive maternal health service delivery (~H) are: (1) efforts of other civil society organisations engaged in the maternal health sector of Kabale; and (2) national reforms to improve maternal health service delivery trickling down to the district.

Operationalising the causal mechanism

This section marks the second step in the theory-testing process tracing procedure where the causal mechanism is explained and made operational by predicting case-specific observable manifestations along with relevant empirical tests to facilitate the drawing of inferences about the found evidence in the case. For illustrative purposes, we operationalise part one (relating to technocrats and female councillors) and part two (relating to female councillors only) at district level respectively. We theorise the necessary scope conditions for these two parts to operate as political will and sensitivity of technocrats and female councillors towards gender equality.

Part 1. We hypothesise that female councillors and technocrats at district level *acquire* knowledge of GRB techniques and *perceive* it as useful to help them implement their respective mandates (h_1). According to evidence from existing GRB evaluations conducted worldwide (UN Women, 2010; Combaz 2013) GRB initiatives have generally increased awareness of gender issues in budgets and built capacities of technocrats and elected representatives to implement GRB. On the basis of this evidence, our confidence in this hypothesised part being true before examining new evidence (prior probability – $p(h_1)$) is relatively high.

One source of evidence to verify h_1 would be interviews with technocrats from the district health department, planning and budget offices, as well as female councillors who participated in the training and technical assistance provided by FOWODE. If h_1 is true, we would expect to obtain interview statements from technocrats and councillors demonstrating an understanding of GRB techniques, along with attestations to the relevance and applicability of GRB techniques in their work. Finding this piece of evidence in the case is critical (certain) for h_1 to hold but not unique, as we might expect that some politicians and technocrats with wide knowledge and exposure will report positively in the interview and attest to what is not the reality. This constitutes a hoop test.

A second source of evidence to verify h_1 would be a review of the FOWODE GRB training manual. We would expect to find coherence between the content of the GRB training manual and interview statements by technocrats and councillors explaining their understanding of GRB principles. We posit that participants who are able to correctly recall and explain the GRB techniques are more likely to have perceived them as useful and

probably implemented them in their work. Finding this evidence is both certain and unique, making a doubly-decisive test.

A third source of evidence to verify h_1 would be reviews of GRB training reports and participant evaluations of the GRB training. We would expect to see account evidence that the GRB training and technical assistance took place in the district, along with the details of the participants who attended. We would also expect to see positive ratings of the quality, relevance and applicability of GRB by technocrats and female councillors from participant evaluations and training reports. Finding this piece of evidence is certain but not very unique. While we might expect that evaluations made immediately after the training are more likely to reflect true opinions of the participants, some participants, as well as the trainer, may be biased towards giving positive feedback about the course that does not reflect reality. This constitutes a hoop test.

Part 2. We hypothesise that female councillors utilise the acquired GRB skills in their oversight of district health policies and budgets for gender-responsiveness (h_2). Existing literature on this hypothesis is mixed: on the one hand, some authors suggest that the

legislature generally lack effective powers to make changes to budgets and policies (Budlender, 2002; Sharp, 2003) while, on the other hand, a few studies (Elson and Sharp, 2010; Combaz, 2013) report positive results. Our prior probability for this hypothesised part ($p(h_2)$) is, thus, conservatively set to be relatively low. This implies that our confidence in the existence of this part of the mechanism will be greatly updated if we can find strong evidence in the case.

One source of evidence to verify h_2 would be interviews with female councillors who participated in the FOWODE GRB trainings. If h_2 is true, we would expect interview statements from female councillors attesting to the application of GRB principles in the evaluation of health department plans and budgets prior to final approval by the district council. Finding this evidence is certain but not unique, as we might expect some female councillors to make false claims that do not reflect reality. This constitutes a hoop test.

A second source of evidence to verify h_2 would be records of proceedings of district council meetings convened to discuss the health department plans and budgets. If h_2 is true, we would expect to find evidence of queries being raised by female councillors about how specific health policies and budgets address gender inequalities, and specifically maternal

health challenges. Finding this evidence is not very certain, as Council minutes may not be easily accessible. However, if the evidence is found, it would be unique to the hypothesised part of the mechanism, resulting in a smoking gun test.

A third source of evidence to verify h_2 would be a review of district health policies and budgets. We would expect to see traces of gender-related health needs being addressed following the GRB training. From a comparison of draft health budgets with approved health budgets on an annual basis since 2011[1], we would expect to see evidence of adjustments being made to budget allocations to address gender-related health needs, more specifically maternal health. Finding this evidence is not very critical for h_2 to hold true, as female councillors could have raised health-related gender concerns but resolutions to change budgets were deferred to later financial years owing to limited funding. If the evidence is found, however, it would strongly confirm that gender inequality queries were raised and acted upon, hence a smoking gun test.

Evaluating the Mechanism- Main Findings

This section marks the third and final step in the theory-testing process tracing procedure. Guided by the predicted empirical evidence in the preceding step, this section presents the

main findings from the case for the two hypothesised parts of the mechanism and applies Bayesian logic to make inferences about our updated confidence in the presence of the specific part of the mechanism linking FOWODE's GRB initiative to improved and gender-responsive maternal health service delivery in Kabale District.

Part 1. Interview statements made by technocrats from the health department, planning and budget offices, community development office, as well as female councillors confirm that FOWODE conducted GRB trainings for selected district-level stakeholders between 2000 and 2013. All interviewees who had participated in the trainings affirmed that the trainings were well delivered, relevant and applicable to their respective duties. Although the trainings had been conducted more than three years ago, the technocrats and female councillors interviewed were able to explain key aspects of GRB principles relating to gender analysis of sectors and mainstreaming gender into policies, plans and budgets. Three out of the six technocrats interviewed attested to have received additional GRB training that was facilitated by the Ministry of Local Government and the Ministry of Gender Labour and Social Development. The found evidence indicates the passing of the hoop test. Passing this hoop test, however, is not sufficient for us to update our confidence in the hypothesised part of the mechanism being true, given that the evidence is not unique to the mechanism.

We would need to evaluate other pieces of evidence to strengthen our confidence in this part of the mechanism.

A review of the content of the FOWODE Gender Budget Manual triangulated the interview statements made regarding the understanding of GRB principles and techniques. The technocrats and councillors interviewed were able to recall and satisfactorily explain at least three out of the five course modules that relate to gender concepts, the budget process cycle and mainstreaming gender in the prioritisation of interventions and allocation of resources. This evidence strongly confirms the hypothesis that technocrats and female councillors acquired knowledge of GRB and perceived it to be useful. The doubly-decisive test passes.

While original participant evaluation forms could not be readily obtained for the GRB trainings, given that it was conducted by external consultants, a review of the training reports and FOWODE periodic reports confirmed that technocrats and female councillors were trained on GRB techniques. From the year 2000 to 2006, FOWODE concentrated efforts on building capacities of district and sub county technocrats to conduct gender analyses of sectors and allocate resources equitably from a gender perspective. District

technocrats also received technical support in the production of a district gender analysis of sectors in 2004 and a district gender policy in 2009. Between 2006 and 2012, female councillors received three trainings on effective legislative engagement with a focus on how to analyse proposed legislation, budgets and policies from a gender perspective and how to prepare motions and lobby for women issues at district council sessions. The ratings of the trainings according to the training reports were positive, showing consistency with the interview statements made. The found evidence passes the hoop test.

Given that we found evidence that is consistent with the two hoop tests and one doubly-decisive test for this part of the mechanism, we can reasonably confirm that this part of the mechanism was present and functioned as predicted. Limited updating, however, has been made to our confidence in the hypothesis being true in the light of the new evidence, given that existing studies had already largely predicted this hypothesis to be true.

Part 2. Interview statements made by female councillors attested to the application of GRB techniques in the evaluation of health policies, plans and budgets prepared by technocrats. The female councillors interviewed revealed that following the FOWODE training, 15 out of the 18 female district councillors formed a women's caucus to periodically monitor

health facilities and identify issues that affect women which they would then collectively prioritise and lobby for at district council meetings. The found evidence passes the hoop test but is not sufficient to confirm the presence of this part of the mechanism.

The minutes of district council health committee meetings could not be readily obtained so the smoking gun test failed. A review of the minutes of the women's caucus meetings, however, revealed details of the health facilities monitored and the issues identified and prioritised, which included staffing of health centres that offer maternal health care with midwives, construction of placenta pits at health facilities and improving access to maternal health facilities by rehabilitating dilapidated roads and bridges. A further triangulation of this observation with interview statements made by male members of the district council health committee affirmed that women caucus members were vocal about the consideration of women's health needs in health plans and budgets. Although the smoking gun test failed, the found evidence points towards the women councillors having utilised the acquired GRB skills in their oversight role. According to Bayesian logic, the failure of the smoking gun test does not permit us to disconfirm the presence of this part of the mechanism.

A review of the annual district budgets revealed that from the year 2011/12 to 2014/15, the district has prioritised the construction of placenta pits and staffing of midwives at selected health Centres II using locally generated revenue, in a bid to address maternal health issues in the district. According to Uganda's health service delivery structure, health centres II are not mandated to provide maternity services. Kabale District Council, however, passed a resolution to equip health centres II located in hard-to-reach villages with maternal health equipment, following a proposal presented by the district council health committee. This constitutes strong evidence for the smoking gun test.

The passing of the hoop test and one smoking gun test enable us to strongly confirm that this hypothesised part of the mechanism (h_2) was present and functioned as predicted. The found evidence also greatly updates our confidence in h_2 being true given that our prior probability (p (h_2)) was low.

The scope conditions. An evaluation of the theorised scope condition for the two parts of the mechanism ascertained that there was political will among the technocrats and councillors towards gender equality, although, not to the full extent. This is evidenced by the commitment to attend the GRB trainings followed by implementation of the acquired skills. Kabale district, however, still lacks an approved gender policy despite having been

supported by FOWODE to develop one in 2009. Whereas the absence of a district gender policy did not affect the functioning of the two parts of the mechanism, it could pose sustainability challenges to the implementation of a gender perspective in district health policies and budgets in the long term.

Conclusion

This article has argued for and demonstrated the application of TBE and PTM in the evaluation of complex interventions such as the effects of civil society gender budget initiatives on maternal health service delivery at local government level. Whereas the last two decades have seen the popularisation of GRB as an effective tool for governments to meet gender equality commitments, limited research has been conducted to evaluate the effectiveness and impact of GRB initiatives worldwide. The limited evidence base has largely been attributed to the complexity and diversity of GRB initiatives. This article has proposed the integration of TBE with PTM to empirically evaluate and make strong within-case causal inferences about the effectiveness and impact of complex GRB initiatives.

The integrated approach starts with TBE to explicate the programme theory of how a particular GRB initiative operating in a given context is expected to influence the intended

outcome. On the basis of the programme theory, PTM explicitly theorises the causal mechanism linking the interventions of the GRB initiative to the intended or observed effects. In doing so, PTM enables us to empirically study the transmission of causal forces from the cause (the GRB intervention) to the intended effects. PTM also provides clear guidance on the types of evidence needed to evaluate the presence of the mechanism along with relevant criteria to judge the strength of the evidence found in a particular case using Bayesian logic. The article has partially illustrated the application of this approach using a case of the FOWODE GRB initiative in maternal health in Kabale District, rural Uganda.

We maintain that the use of the TBE-PTM approach in an in-depth case study is particularly valuable for evaluating gender mainstreaming interventions as it explicitly elaborates the activities, processes and mechanisms through which interventions systematically bring about political, social and cultural transformation. We conclude by emphasising that the TBE-PTM approach facilitates better programming of gender mainstreaming interventions for enhanced success.

Acknowledgements
Appreciation goes to Dr. Henry Manyire at the School of Women and Gender, Makerere University Kampala Uganda, for the technical support during field research.

Funding

This work was supported by the Flemish Interuniversity Council (VLIR-UOS) [ICP PhD 2012-010].

Notes

1. The five-year political term for the female councillors interviewed for the study commenced in 2011 and ends in 2016.

References

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

Befani B and Mayne J (2014) Process Tracing and Contribution Analysis: A Combined Approach to Generative Causal Inference. *IDS Bulletin,* 45: 17-36.

Bennet A (2014) Appendix: Disciplining our conjectures: Systematising process tracing with Bayesian analysis. In Checkel JT and Bennet A (eds) *Process Tracing: From Metaphor to Analytic Tool.* New York: Cambridge University Press, pp. 276-298.

Birckmayer J and Weiss C (2000) Theory-based evaluation in practice: what do we learn?. *Evaluation Review,* 24(4): 407-431.

Budlender D (2002) A global assessment of gender responsive budget initiatives. In Budlender D, Elson D, Hewitt, G and Mukhopadhyay T (eds) *Gender Budgets make Cents: Understanding Gender Responsive Budgets.* London: The Commonwealth Secretariat, pp. 83-130.

Budlender D (2009) Integrating gender-responsive budgeting into the aid effectiveness agenda. Ten-country overview report, New York: United Nations Development Fund for Women.

Budlender D and Hewitt G (2006) *Engendering Budgets: A Practitioner's Guide to Understanding and Implementing Gender-Responsive Budgets.* London: The Commonwealth Secretariat.

Carlitz R (2013) Improving transparency and accountability in the budget process. *Development Policy Review,* 31(S1): s49-s67.

Checkel JT and Bennet A (2014) Beyond metaphors: standards, theory and the 'what next' for process tracing. In: Bennet A and Checkel JT (eds) *Process Tracing: From Metaphor to Analytic Tool.* New York: Cambridge University Press, pp. 260-275.

Chen H (1990) *Theory Driven Evaluations.* Thousand Oaks, CA: Sage.

Collier D (2011) Understanding process tracing. *Political Science and Politics,* 44(4): 823-830.

Combaz E (2013) Impact of gender-responsive budgeting. Available at: http://gsdrc.org/docs/open/HDQ977.pdf. (accessed 16 November 2015).

Delahais T and Toulemonde J (2012) Applying contribution analysis: lessons from five years of practice. *Evaluation* 18(3): 281-293.

Elson D (2002) Gender responsive budget initiatives: Key dimensions and practical examples. In Judd K (ed) *Gender Budget Initiatives: Strategies, Concepts and Experiences*. New York: United Nations Development Fund for Women, pp. 15-29.

Elson D and Sharp R (2010) Gender-responsive budgeting and women's poverty. In Chant S(ed) *International Handbook on Gender and Poverty: Concepts, Research, Policy*. Cheltenham: Edward Elgar Publishing Ltd, pp. 522-527.

George AL and Bennet A (2005) *Case Studies and Theory Development in the Social Sciences.* Massachusets: Massachusets Institute of Technology Press.

Goetz AM and Jenkins R (1999) Accountability to Women in Development Spending - Experiments in Service-Delivery Audits at Local Level. Paper Presented at UNDP Conference on Gender, Poverty and Environment-Sensitive Analysis, New York.

Hickey G, McGilloway S, O'Brien M, Leckey Y and Delvin M (2015) A theory-based evaluation of a community-based funding scheme in a disadvantaged surbarban city area. *Evaluation and Program Planning* 52: 61-69.

Hofbauer H and Garza M (2009) The missing link: applied budget work as a tool to hold governments accountable for maternal mortality reduction commitments. Report, International Budget Partnership and the International Initiative on Maternal Mortality and Human Rights, May.

Jacobs B, Ir P, Bigdeli M, Annear P, and Van Damme W (2012) Addressing access barriers to health services: An analytical framework for selecting appropriate interventions in low-income Asian countries. *Health Policy and Planning* 27(4): 288-300.

Jahan R (1996) The elusive agenda: mainstreaming women in development. *The Pakistan Development Review* 35(4): 825-834.

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

Mayne J (2012) Contribution analysis: Coming of age. *Evaluation,* 18(3): 270-280.

McGee R and Gaventa J (2010) Review of Impact and Effectiveness of Transparency and Accountability Initiatives: Synthesis Report Brighton: Institute of Development Studies.

Papp S, Gogoi A and Campbell C (2013) Improving maternal health through social accountability: A Case study from Orissa, India. *Global Public Health* 8(4): 449-464.

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

Robinson M (2008) Improving transparency and promoting accountability. In Robinson M (ed) *Budgeting for the Poor.* Basingstoke: Palgrave Macmillan.

Rogers P, Petrosino A, Huebner T and Hacsi T (2000) Program theory evaluation: Practice, promise and problems. *New Directions for Evaluation* 2000(87): 5-13.

Rohlfing I (2014) Comparative hypothesis testing via process tracing. *Sociological Methods and Research,* 43(4): 606-642.

Schmitt J and Beach D (2015) The contribution of process tracing to theory-based evaluations of complex aid instruments. *Evaluation,* 21(4): 429-447.

Scriven M (2008) A summative evaluation of RCT methodology: An alternative approach to causal research. *Journal of Multidiciplinary Evaluation,* 5(9): 11-24.

Sharp R (2003) *Budgeting for Equity: Gender Budget Initiatives within a Framework of Performance Oriented Budgeting.* New York: United Nations Development Fund for Women.

Sharp R (2007) Gender responsive budgets (GRB's) have a place in financing gender equality and women's empowerment. Available at: http://www.un.org/womenwatch/daw/egm/financing_gender_equality/ExpertPapers/EP.4%20Sharp.pdf. (accessed 16 November 2015).

Sharp R and Elson D (2012) Improving budgets: a framework for assessing gender responsive budget initiatives. Available at: http://www.unisa.edu.au/Documents/EASS/HRI/gender-budgets/sharp-elson-improving-budgets.pdf. (accessed 16 November 2015).

Sodani P and Sharma S (2008) Gender Responsive Budgeting. *Journal of Health Management* 2(10): 227-240.

Stern E, Stame N, Mayne J, Forss K, Davies R and Befani B (2012) Broadening the Range of Designs and Methods for Impact Evaluations. DFID working paper 38, Department for International Development, April.

Suchman E (1967) *Evaluative Research.* NewYork: Russel Sage Foundation.

United Nations Population Fund and United Nations Development Fund for Women (2006) *Gender Responsive Budgeting and Women's Reproductive Rights: A Resource Pack.* New York: United Nations Population Fund and United Nations Development Fund for Women.

Van Eerdewijk A and Brouwers J (2014) Gender and theories of change. Report, 4th e-discussion Endnote, Hivos, June.

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

Weiss C (1972) *Evaluation Research: Methods for Assessing Program Effectiveness.* Englewood Cliffs, NJ: Prentice Hall.

Weiss C (1995) Nothing as practical as a good theory: exploring theory-based evaluation for comprehensive community initiatives for children and families. In Connell JP,

Kubisch AC, Schorr LB and Weiss CH (eds) *New approaches to evaluating community initiatives*. Washington, DC: Aspen Institute, pp 65-69.

Weiss C (1997) Theory-based Evaluation: Past, present and future. *New Directions for Evaluation* 76: 41-55.

White H (2000) Theory-based impact evaluation: principles and practice. *Journal of Development Effectiveness* 1(3): 271-284.

White H and Phillips D (2012) Addressing Attribution of Cause and Effect in Small n Impact Evaluations: Towards an Integrated Framework. Working paper, International Initiative for Impact Evaluation.

UN Women (2010) Evaluation report, UNIFEM's work on gender responsive budgeting: Overview. Report of the Evaluation Unit 2009, United Nations Development Fund for Women.

World Bank (2003) *World Development Report 2004: Making Services Work for Poor People.* Washington, DC: The International Bank for Reconstruction and Development/The World Bank.

World Bank. (2005). Social Accountability: What does it mean for World Bank? *Social Accountability Sourcebook.* Available at: http://www.worldbank.org/socialaccountability_sourcebook/PrintVersions/Conceptual%2006.22.07.pdf. (accessed at 16 November 2015).

WHO (2014) *World Health Statistics 2014: A Wealth of Information on Global Public Health*. Geneva: World Health Organisation.

Table 1. Types of empirical tests and related attributes.

Source: adapted from Rohlfing (2014: 610)

Test type	**Description**
Doubly-decisive	Empirical predictions combine high uniqueness and high certainty Passing the test is both necessary and sufficient for inferring causation
Smoking gun	Empirical predictions have high uniqueness, but low certainty Passing the test is sufficient but not necessary for inferring causation
Hoop	Empirical predictions have low uniqueness, but high certainty Passing the test is necessary but not sufficient for inferring causation
Straw-in-the-wind	Empirical predictions combine low uniqueness and low certainty Passing the test is neither sufficient nor necessary for inferring causation

Figure 1. Programme theory for local-level civil society gender budget initiatives in maternal health.

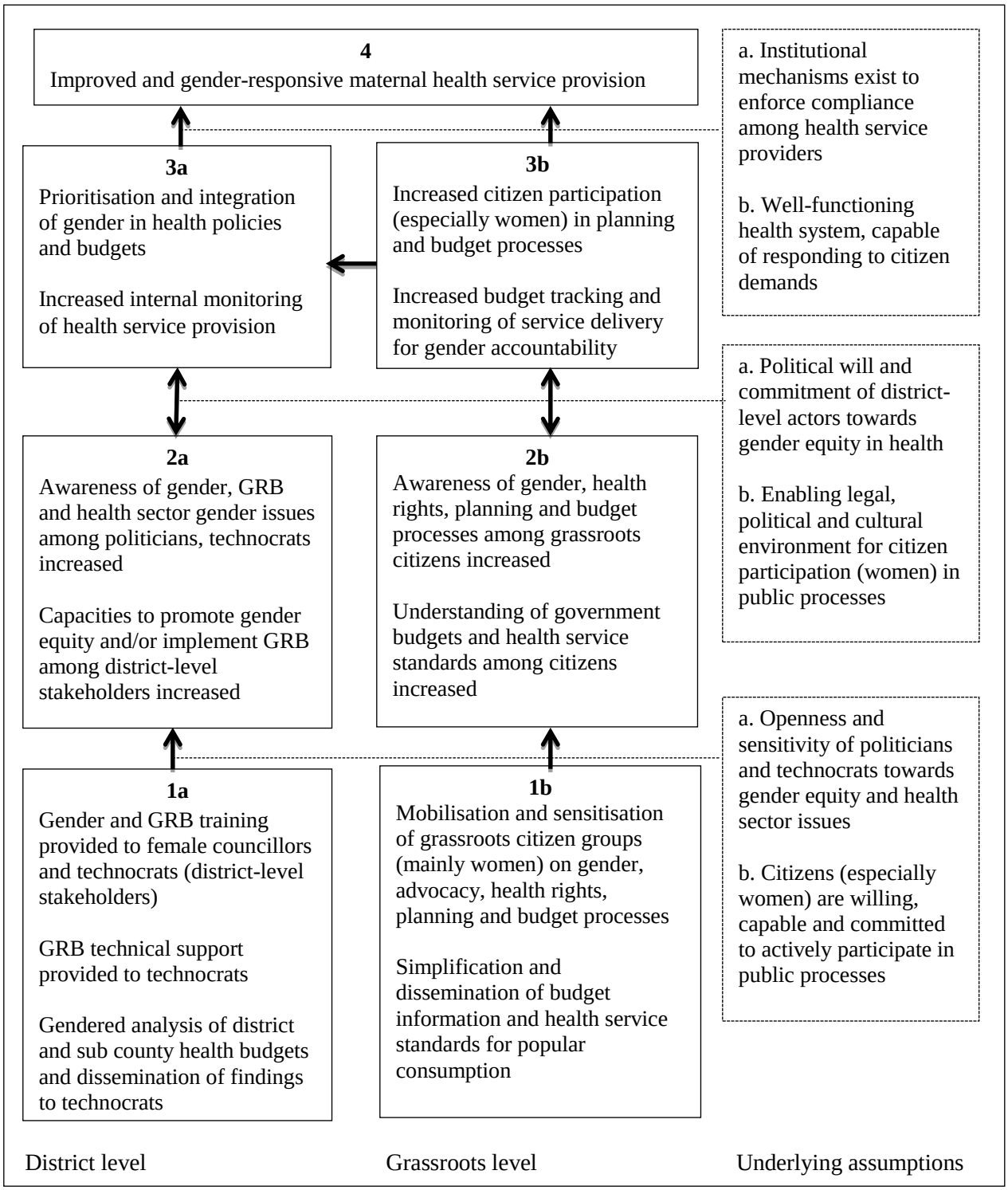
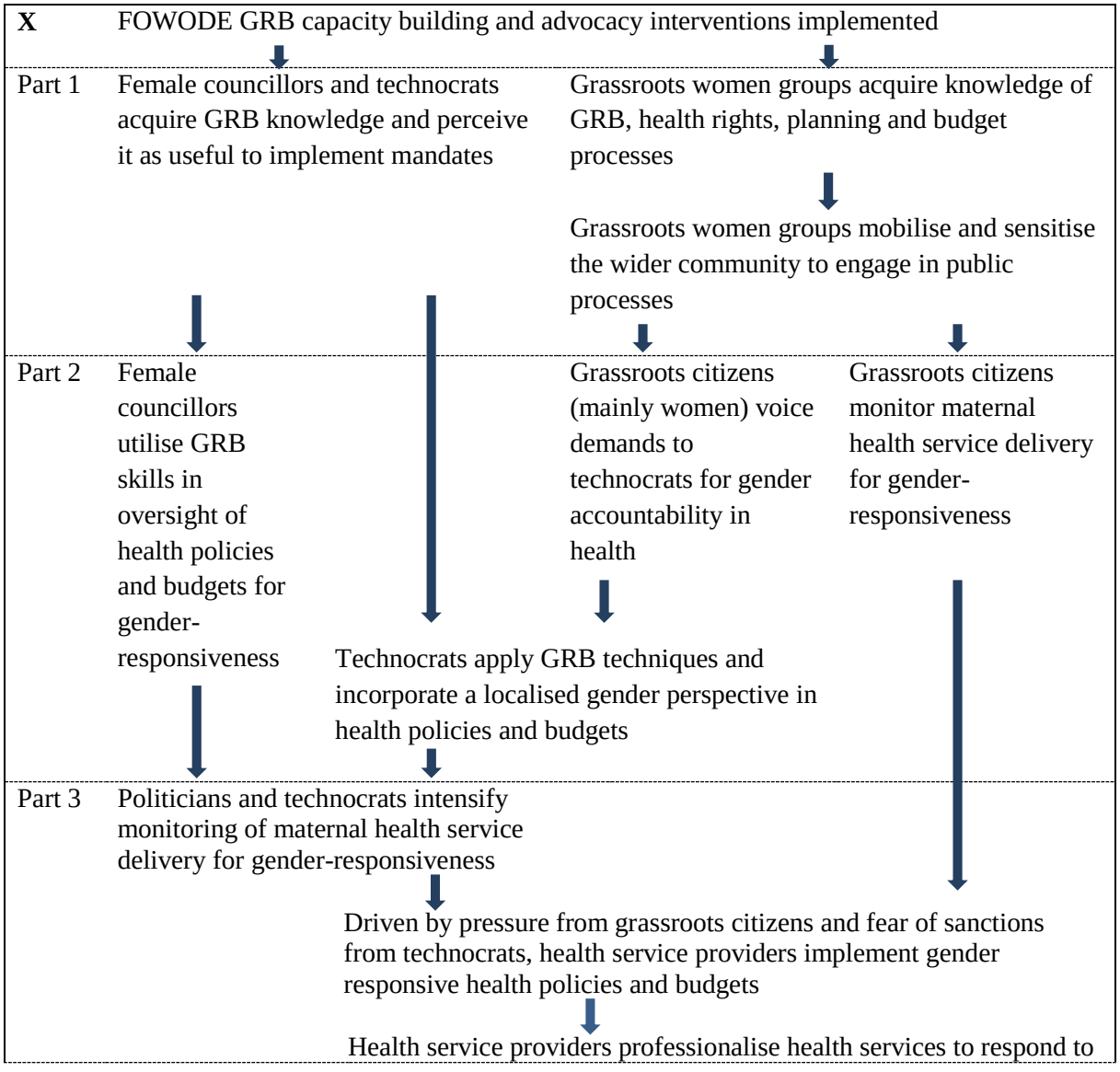


Figure 2. Causal mechanism for FOWODE's GRB initiative in maternal health in Kabale District



	citizen needs ⬇
Y	Improved and gender-responsive maternal health service delivery