Community advocacy forums and public service delivery - Impact, and the role of information, deliberation, and administrative placement

Nassul Kabunga*, Caroline Miehe†, Tewodaj Mogues‡, Bjorn Van Campenhout§†

December 26, 2019

Abstract

To improve governance and public service delivery public service delivery, the Government of Uganda organizes community forums – popularly known as barazas – where citizens receive information from government officials, and get the opportunity to directly engage with them. We run a cluster randomize control trial to assess the impact of the baraza intervention on a range of outcomes related to agriculture, health, education and infrastructure. Using a factorial design, we further test the relative importance of the two main components of the intervention – information provision and citizen engagement. Furthermore, we compare the effectiveness of barazas organized at the district level to the effectiveness of barazas organized at the sub-county level, as the administrative placement of the barazas is a key determinant of the cost-effectiveness of this policy intervention.

1 Introduction

In 2015, we designed a study aimed at evaluating the effectiveness of community advocacy forums, also known as barazas, in Uganda. The baraza programme, an initiative of the president of Uganda and implemented by the Office of the Prime Minister (OPM), was designed to improve public service delivery by enhancing public involvement in holding the government accountable for service delivery in relation to the resources spent. The study had several objectives.

*Development Strategy and Governance Division, International Food Policy Research Institute, Kampala, Uganda

†LICOS Centre for Institutions and Economic Performance, KULeuven, Leuven, Belgium

‡International Monetary Fund, Expenditure Policy Division, Fiscal Affairs Department, Washington DC, USA

§Development Strategy and Governance Division, International Food Policy Research Institute, Leuven, Belgium - corresponding author: b.vancampenhout@cgiar.org

First, it wanted to establish, in a rigorous way, if the program had an impact on public service delivery. Second, the study proposed to compare the effectiveness of barazas organized at lower administrative levels (the subcounty) to that of barazas that are organized at a more aggregate level (the district), as the level of administrative placement is an important determinant of the cost-effectiveness of the policy intervention. Third, the project also aspired to explore pathways through which community advocacy forums may affect outcomes. In particular, using a two-by-two factorial design, it differentiates between the impact of (1) providing citizens with information related to budgeting and planing, and (2) the impact of letting citizens engage with public servants and politicians in a questions-and-answers session. The project was funded by the International Institute for Impact Evaluations (3ie) after a competitive call for proposals, and a baseline survey involving more than 12,500 households and 400 government officials was conducted. After completion of the baseline, we trained local government officials to ensure adherence to the intervention protocols, and the interventions were rolled out by the OPM, our main implementing partner.

While the project was initially assumed to take about 2 years, OPM faced various complications that affected the timely roll-out of the barazas, including budgetary constraints and disruptions related to the general elections of 2016. Four years after the baseline survey, with about 50 percent of the planned subcounty barazas implemented, a trade-off needed to be made between waiting for the remaining barazas to be completed or conducting the endline after partial roll-out. It was decided to proceed with endline data collection and employ estimation and data collection strategies to control for potential selection bias that may have been introduced due to the partial roll-out.

The current version of this paper serves as a registered report. In particular, it was prepared as a "mock report", which contains the analysis on simulated endline data for a core set of primary outcomes to test the four primary research questions (impact of sub-county barazas, relative importance of information component, relative importance of deliberative component, and importance of administrative placement). The outcomes can be categorized into four broad sectors: agriculture, health, education and infrastructure (including drinking water and roads). With this report, we thus commit to key outcomes in each sector, and combine the outcomes in the four sectors into sector level indices and in a single index following Anderson (2008), to account for multiple hypothesis testing. Once endline data has been collected, we will simply replace the simulated data with real data, and we will immediately be able to publish the report by running the pre-coded analysis. Pre-registration and mock reports are effective tools against fishing and false-positive science (Humphreys, De la Sierra, and Van der Windt, 2013).[1]

There have been several studies that look at the impact of community involvement on public service delivery, many of then done in Uganda. A landmark

[1] This document was prepared using Lyx, an open source Latex front-end. All Latex and R code to replicate the analysis is placed under revision control using Git. The R code is automatically executed when the Lyx document is compiled (using the R package knitr). The Git repository can be found at https://github.com/bjvca/baraza/.

study is Björkman and Svensson (2009) that looks at the impact of a community driven local accountability project in primary health care provision in Uganda. They find that the intervention resulted in significant improvements in health care delivery, utilization, and health outcomes (e.g., most notably child mortality) after one year, and confirm in Björkman Nyqvist, de Walque, and Svensson (2017) that these effects are still present more than four years after the initial intervention despite minimal follow-up. More recently, however, Raffler, Posner, and Parkerson (2018) come to more nuanced conclusions when testing an intervention closely modeled on the one of Björkman and Svensson (2009) in a similar setting and implemented by two NGOs. The study, involving a three wave panel of more than 14,000 households and a factorial design to break down the intervention into its two most important components similar to what we use, validates the power of information provision to change the behavior of front-line service providers, but casts doubt on the ability to foster community monitoring or to generate improvements in health outcomes, at least in the short term.

Our study contributes to this literature in various ways. First, this study is one of the few that considers the role of administrative placement on the effectiveness of community monitoring. The level at which the intervention occurs may affect its effectiveness in opposing ways. On the one hand, interventions at a more local level may result in more relevant issues being scrutinized. However, qualitative explorations suggest that often, issues raised in lower level barazas fall under the responsibility of higher levels of government or other institutions that are beyond the operational jurisdiction of the participating officials (Van Campenhout et al., 2018). This may be less of a problem when barazas are organized at district level. Most other studies consider interventions that are placed at fairly local levels. For instance, the intervention in Raffler, Posner, and Parkerson (2018) is implemented in health centers and their associated catchment areas.

Second, our study evaluates the impact of a government initiative, which may instigate an entirely different set of dynamics than interventions that are organized by local or international NGO's. It has been argued that successful devolution can only happen in the context of a strong state, able to ensure consistent regulation, and a well-informed public backed up by a participatory political culture (Golooba-Mutebi, 2005). Many of the actors involved may find that NGOs are not mandated when it comes to public services such as health and education. Furthermore, also here, it may be easier to re-allocate resources to problems identified during barazas is it is organized by the government. However, effects may also work in the opposite direction. For example, an intervention to reduce absenteeism in government public health facilities in India was initially very successful, but ceased to have any impact after the local bureaucracy started providing official excuses for most of the nurses' absences (Banerjee, Duflo, and Glennerster, 2008). Most other studies work with NGOs (Björkman and Svensson, 2009; Raffler, Posner, and Parkerson, 2018)

Third, baraza's take a comprehensive, multi-sector approach, enabling cross-sectoral planning and potentially allowing for re-allocations across sectors. Some of the problems most mentioned by users, such hygiene in health centers or

accessibility, involve cooperation between heads of different sectors (eg health and infrastructure to get access to water in health centers). Bringing sector heads together and confronting them with the demands of citizens may increase information sharing and cooperation between them. Most other studies focus on a single sector; the health sector in particular seems to be a popular sector for community monitoring interventions (Arkedis et al., 2019; Björkman and Svensson, 2009; Raffler, Posner, and Parkerson, 2018)

In this report, we start by providing a brief overview of the government program we study. We motivate the main research questions in a separate section. We then present the design of the original cluster randomized control trial and use baseline information to assess balance. We then discuss implementation challenges, and explain strategies we will use to address these. A next section presents updated power calculations. Results are presented based on simulated endline data drawn from the baseline, merely to show what the final analysis would look like. In a final section we outline how endline data will also be used to further assess the likelihood that the implementation challenges introduced selection bias.

2 Background of the baraza impact evaluation

Since Uganda's independence in 1962, Uganda's development efforts have been thwarted by political turmoil and economic mismanagement. In the mid-1980s, after attaining relative stability, the Government of Uganda, supported by development partners, initiated far reaching liberalization efforts and introduced a decentralized system of governance (Francis and James, 2003). Decentralization was particularly viewed as a suitable mechanism for improving efficiency, effectiveness and inclusiveness of public service delivery, for the formulation of services that are more aligned to citizen's needs, for bringing representative governance closer to citizens, and for the empowerment of local stakeholders. However, in order to work, decentralization requires local citizens to actively participate in planning, implementation, monitoring and evaluation of development interventions in their locations, so as to improve accountability and responsiveness of local leaders and service providers.

Until recently, this empowerment component has been largely lacking in Uganda's decentralization process. The realization of benefits of decentralization in Uganda has been greatly affected by ineffective monitoring and weak accountability mechanisms, especially with respect to beneficiaries holding the service providers accountable (Björkman and Svensson, 2009; Reinikka and Svensson, 2004). In response to this, the Government of Uganda, under the stewardship of the Office of the Prime Minister (OPM), initiated community advocacy forums (or citizen barazas) in 2009 with the general objective of "enhancing public involvement in holding the government accountable for service delivery in relation to the resources spent" (OPM, 2013).

Barazas are platforms for enhancing information sharing between policy makers, public servants, and beneficiaries of public goods and services. In ad-

dition, it provides the opportunity for citizens to ask questions to their leaders and deliberate among themselves. With barazas, citizens in particular have the opportunity to participate in the policy process by directly engaging with service providers, and to demand accountability of the use of public resources. It is expected that, ultimately, barazas will contribute to effective monitoring, and increase accountability and transparency among all stakeholders.

Barazas have been implemented in Uganda for about 10 years by now. Barazas were first piloted in the financial year 2009/10.[2] Since then, efforts have been underway to roll out barazas in all subcounties in the country. During the full-scale implementation phase in the financial year 2010/2011, 16 more subcounties in 8 districts had held a baraza meeting. And, by the last quarter of 2011/2012, 267 out of the country's total of 1,340 subcounties, spread in 112 districts had held a baraza meeting. At the beginning of 2012/2013 however, changes in implementation were suggested: subsequent barazas would target district-level reporting so as to increase participation at a higher level.

A typical baraza is initiated from the center, with the Office of the Prime Minister mobilizing district and subcounty officials. These include the Chief Administrative Officer (CAO) as the head of public service delivery at the district level, the Resident District Commissioner as a direct representative of the president (RDC), the District Local Council Chairperson (LC5) as the representative of political leadership at the district level, and the various sector heads. Especially for barazas organized at the subcounty level, the subcounty level equivalents of the CAO (the subcounty chief) and the LC5, the subcounty chairperson (LC3) also have important roles. OPM, in consultation with the district leaders (RDC, CAO and the LC5) and other stakeholders, agree on the date and a neutral venue in which to hold the baraza event. Again, in consultation with the district leaders, a viable moderator and an interpreter into the local language where applicable are identified to guide the baraza forum. Village mobilizers and community resource persons are used to publicize the event. These community mobilization efforts are further reinforced by adverts in the local media in the form of radio announcements; printed banners, posters and fliers, and mobile public address systems, a few days before the baraza event.

A baraza meeting is chaired by the Office of the RDC in each district. In front of the audience, including local citizens and invited opinion leaders and elders, the RDC seeks accountability and feedback from each head of major sectors: health, education, physical infrastructure (mainly water facilities and roads), and agriculture (particularly national agricultural advisory services). Sector heads are required to present what services were planned to be delivered in the subcounty; what was actually delivered and in what quantity and quality; and what issues and challenges have emerged and what is the way forward. The RDC then seeks reactions and feedback from citizens on whether what has been presented is what was planned for and actually implemented in different

[2]The initial pilot barazas were undertaken in eight lower level local governments (generically referred to as subcounties) of the four districts of Masaka, Bushenyi, Kumi and Nebbi, which are respectively located in the four geographical regions of Uganda: Central region, Western region, Eastern region, and Northern region.

locations. Sector heads are then given another opportunity to clarify on or react to any issues raised by the citizens. At the end of the process, the RDC makes a report to the OPM, indicating issues that arose in the baraza meeting. This report particularly points out policy and program implementation weaknesses and challenges, which is then expected to further feed into the general government performance management system. In general, a minister of state will also be present at the baraza.

3 Research questions: impact, information provision versus deliberative aspect, and administrative placement

3.1 The impact of (subcounty level) barazas

The baraza intervention fundamentally seeks to improve public services through improving downward accountability of local public decision makers and service providers. The baraza intervention as conceived by the OPM is a fairly standard community based monitoring intervention that combines the provision of information with the possibility of citizens to engage with each other and with decision makers at a fairly local level. Such community based monitoring has become a popular tool to increase service delivery. However, not all such interventions appear to be successful (Olken, 2007). A first question is therefore simply related to the impact of a typical baraza intervention as organized by OPM.

3.2 The Information Mechanism

One of the 'political market' imperfections that looms large in public service outcomes for the poor is lack of information on the part of citizens about the actions and performance of public agents (Keefer and Khemani, 2005). Even where democratic institutions are strong and public decision makers are compelled to take into account citizen preferences in their policy and public expenditure decision making, limited access to information by citizens can be a serious hindrance to their policy preferences being realized.

Information imperfections in this context can take two fundamental forms: Lack of information about available publicly provided goods, services, and infrastructure; and lack of information about public agents' efforts (including actions, policies and expenditures). The first type of information constraint can lead residents to fail to take advantage of the services that are in fact available to them, and thus it ultimately brings about poor usage of even extant services. The second type of information constraint makes it difficult for residents to hold public decision makers accountable, leading to potential undersupply of the needed services. When either type of information constraint is, furthermore, more binding among the poor in society than among the non-poor, all else equal,

this results in negative distributional effects of service provision. The hardest type of information deficit to overcome—and potentially the most important for outcomes—is a particular combination of both lack of knowledge about services and about public efforts, namely the challenge for citizens to be aware of the causal link between public actors' efforts and the existing services.

Channeling of information to citizens about the quantity, modality, and quality of public services, as well as about the investments and policy decisions made by politicians, bureaucrats, and service providers can increase the ability of the users to hold the leaders accountable to improve service provision. For example, Pandey, Goyal, and Sundararaman (2009) establish using a field experiment in India that community information campaigns about states' school management obligations had a positive impact on school performance. Gilens (2001) identifies a significant influence of providing policy facts on the public's political judgment. Grossman and Michelitch (2018) disseminate information about job performance for randomly selected Ugandan politicians. While this increases job performance for the polictician increases among a range of criteria, there is no impact on public service provision. A recent review of 48 empirical studies on the impacts of information on governance and service delivery also suggests that the availability of information alone may not suffice. Information must be deemed relevant to its recipient, and individuals must have both the power and incentives to act on the information (Kosec and Wantchekon, 2020).

3.3 The Deliberation Mechanism

Theory on the deliberative process explores the effects of greater both-way interaction between citizens and leaders (versus only one-way information provision, as in the first mechanism described above), and of interaction among citizens with each other. Much of the work on local deliberative processes have as their foundation broader theories of deliberative democracy (eg. Gutmann and Thompson, 1998). Features of deliberative democracy include, among others, that agents are given the opportunity to articulate their diverse positions to each other; that these positions are conscientiously mutually vetted based on the evidence presented in their support; and that an agent's political standing, economic status, or other characteristics reflecting their power is not a factor in the weight placed on their arguments.

There are various ways in which deliberation increase the quality of public service delivery. Firstly, it has a legitimating effect on decisions arrived at in this fashion. These follow from the particular features, as described above, that highlight equal voice of the arguments of both marginal and advantaged agents, and the role of evidence that support the positions articulated. Secondly, it is judged to have the ability to more effectively distill social choice than simple voting and majoritarian rule, in part by building of consensus both among citizens and between public servants and citizens . Thirdly, it is said to have a positive impact on the vigor and breadth of subsequent citizen involvement in community affairs (Björkman Nyqvist, de Walque, and Svensson, 2017). In deliberation with public officials, citizens get to observe how leaders

respond to their queries and expression of dissatisfaction with services, which can have an effect of feeling empowered that they can in fact ask for, and are entitled to, better or more resource allocation to the public services that they need. The perception that one is entitled to and can demand changes in public action is an important precondition for holding public servants accountable to improve services. There are several reasons why this perception on the part of citizens can be low—especially in authoritarian systems but to some extent even in democratic systems. Thus, the fact that this feature is being addressed through the baraza intervention allows testing for its hypothesized effectiveness to improve service delivery. Impacts of deliberative processes have also been the subject of empirical analysis. For example, experimental evidence shows that deliberative processes make decision outcomes less sensitive to the institution (e.g. voting) rules that bring them about (Goeree and Yariv, 2011) or may reduce the prevalence of clientelism (Fujiwara and Wantchekon, 2013).

3.4 Administrative placement

The baraza intervention can be distinguished by the administrative level at which it is implemented: Barazas had been originally planned to be implemented at the subcounty level but in recent years, the focus is being turned to implementing barazas at the district level. This administrative placement dimension immediately points to a potential tradeoff between attempting to achieve breadth of coverage (through the district-level barazas), and attending to depth and quality of coverage (through subcounty-level barazas). While conducting a district-level baraza may be cheaper than conducting subcounty-level barazas in all subcounties of that district, it is not clear a priori how these cost savings justify potential reduction in effectiveness of district-level barazas in any given subcounty of the concerned district. Therefore, another vital aspect of this study will be to compare the effectiveness of barazas organized at the higher district level to those organized that the lower subcounty level.

4 Experimental design and balance at baseline

As noted in Section 2, until now, a small share of all subcounties, albeit located throughout all of Uganda's 112 districts across the four regions of the country, have received an SC-level baraza intervention. We designed a social experiment covering districts, subcounties, and households across the four regional blocks (Northern, Western, Central and Eastern) of Uganda. Each regional block has somewhat unique characteristics in terms of ethnicity, geographical and agro-ecological conditions, as well as cultural history. We make our sample selection of districts from among 'eligible districts', and our sample selection of subcounties from 'eligible subcounties' (or 'eligible SCs'). An 'eligible district' is defined here as a district in which a district-level baraza has not already been implemented prior to this study. The districts with already implemented district-level barazas are excluded from the sampling frame because we cannot track which

members of which subcounty attended the district-level baraza. An 'eligible SC' is defined as a subcounty to which two conditions apply: (i) a subcounty-level baraza has not already taken place, and (ii) the subcounty is not located in a district in which a district-level baraza has already been implemented i.e. the subcounty is located in an eligible district. Preliminary analysis of the baraza implementation data from OPM shows that there are 20 or more eligible districts per region, amounting to a total of 94 eligible districts. In each region, there are at least 147 subcounties that have never been treated as well as are in eligible districts; the total of such eligible subcounties is 722.

As a first step, this study proposes a nested, or two-step, randomization design that will test the effects of two different administrative placement dimensions of the baraza intervention in a nested way. The first dimension has implicitly already been alluded to in various places above: The baraza intervention can be distinguished by the administrative level at which it is implemented: Barazas had been originally planned to be implemented at the subcounty level but in recent years, the focus is being turned to implementing barazas at the district level. This administrative placement dimension immediately points to a potential tradeoff between attempting to achieve breadth of coverage (through the district-level barazas), and attending to depth and quality of coverage (through subcounty-level barazas). While conducting a district-level baraza may be cheaper than conducting subcounty-level barazas in all subcounties of that district, it is not clear a priori how these cost savings justify potential reduction in effectiveness of district-level barazas in any given subcounty of the concerned district. Therefore, this study will be to compare the effectiveness of these two different level administrative placements of the treatment.

The second dimension of the study is directly derived from the theoretical framework and empirical literature discussed earlier in Section 3, covering similar governance interventions in developing countries to enhance the quality of public services and local accountability. As mentioned before, barazas are viewed as platforms for enhancing information sharing between policy makers, development partners and beneficiaries of public goods and services. In addition, barazas provide the opportunity for citizens to ask questions to and go into debate with their leaders, ultimately contributing to effective monitoring, accountability and transparency among all stakeholders. It is useful to know if there is a differential effect between these two major features within the treatment, which may then be useful information to improve the effectiveness of the intervention in the future. To this end, we will appropriately design a randomization experiment that will have untreated subcounties as a control group and treated subcounties where subcounty-level barazas will be conducted. However, instead of just dividing the sample into treatment and control, we will differentiate between two key components that can grossly be identified within the current barazas: (i) the information sharing component and (ii) the deliberative component, as theoretically motivated in Section 3, and described in Section 4 in the specific barazas context as executed by OPM to date. We propose to use a factorial design. If we define two levels within each factor, a 2 by 2 fully crossed factorial design is appropriate and can allow us to describe the baraza

intervention in terms of two separate and one crossed treatment. As such, we can structure the evaluation based on 3 treatment arms in an experiment as follows (a graphic illustration follows in Figure 2 further below):

- Crossed treatment arm (S 0ID):- includes sampled subcounties within the context of the administrative placement dimension (baraza treatments and no-treatments at district-level) explained earlier. This comprises of a subset of treated subcounties that will receive the current form of the baraza as it is currently implemented, with both the information sharing component and the deliberative component (i.e. the crossed treatment). The deliberation treatment arm (S 0D):- is directly derived from the existing modality of the barazas as they are conducted at the subcounty level heretofore, i.e. as described in Section 4 above. However, since this treatment wants to separate out the effect of citizens challenging government officials and deliberating among one another, this treatment will not involve the information sharing components of the existing barazas as described in Section 4. In other words, this treatment will not involve the presentations mentioned, but only involve a facilitated session where citizens can engage with each other and with government officials.

- The deliberation treatment arm (S 0D):- is directly derived from the existing modality of the barazas as they are conducted at the subcounty level heretofore, i.e. as described in Section 4 above. However, since this treatment wants to separate out the effect of citizens challenging government officials and deliberating among one another, this treatment will not involve the information sharing components of the existing barazas as described in Section 4. In other words, this treatment will not involve the presentations mentioned, but only involve a facilitated session where citizens can engage with each other and with government officials.

- The information sharing treatment arm (S 0I):- while S 0D above is focused on the deliberative process, S 0I will concentrate on one-way information provision and sharing. This treatment also follows the model of the pre-existing baraza, but focuses on only the first part that consists of presentations from each head of major sectors on planned activities, on what has been delivered and on the challenges that have emerged. In this treatment, there will be no space for interaction with the government officials or of the community members among each other. It will be seen as an information dissemination event. In other words, this treatment will not involve the facilitated session where citizens can engage with government officials, but only involve the presentations by different sector heads.

The treatment arms described, will examine in absolute terms the effectiveness of information provision, and the deliberative process and the interaction effect thereof, as well as will allow a relative comparison of these two avenues of improving accountability of service providers to the service users. Moreover, such a crossed design allows us to learn which aspects of the baraza is most effective. For instance, not only does this allow us to learn how much more (or less)

effective barazas are compared to the control, but also how much more effective barazas are compared to a treatment that consists only of the information sharing component, or only of the deliberation component. We will operationalize this examination in our experimental design after more detailed discussions with OPM on the above proposed partition of the baraza in the two treatment arms, and based on these discussions, more detailed specifications of the two intervention elements.

While comparing control areas to subcounties that received treatment S 0D , treatment S 0I or both is of main interest, it may also be instructive to compare each treatment to the crossed treatment. For example, it may be that sector heads, through disseminating information, are able to influence the deliberative process. Sector heads may be able to play down expectations of citizens by presenting unambitious plans and exaggerating delivered services. This may result in the communities being far less critical and a muted deliberative process may result, as compared to a treatment with only a deliberative component. On the other hand, in a treatment with only a deliberative process, expectations of citizens may be completely unaligned to planned activities, leading to an unproductive deliberative process.

The research proposal contained an extensive series of power calculations that used data from the Uganda National Household Survey of 2009/10 and the Demographic and Health Survey of 2011 to estimate standard errors of the outcomes and interclass correlations. Outcomes used to determine sample size included weight-for-age z-scores for children; number of days unable to work as percentage of days sick at the household level; number of years the average child within the household goes to school, % children in household currently attending school; the proportion of households that was visited by an extension worker in the previous year; maize yields; time to get drinking water (incl. waiting time; and share of households having access to improved drinking water sources. This resulted in the selection of a total sample size of 11,500 households distributed over 230 subcounties throughout Uganda, on which baseline data was collected[3]. More details on the power calculations can be found in the original proposal, which is available as an online appendix. In this document, we will run a series of updated power calculations to account for the implementation challenges (see Section 5)

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
```

[3]We added an additional 3 subcounties in each of the five treatment groups to account for attrition.

```
## [1] 9
## [1] 10
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
```

In Table 1, we test for balance between the treatment groups at baseline following the initial design of the experiment. During baseline, information on 12,545 households was collected. We see that the average household consists of about 6 household members (first column). When comparing households that were exposed to a sub-county level information baraza to households that did not receive a sub-county level information baraza, we see that that households are slightly larger in the treatment group, and the difference is significant at the 5 percent level (second column). The average household has two to three children attending a public school. We also find a slight pre-treatment imbalance on this outcome for the information treatment, but the difference is only significant at the 10 percent level.

In the third column of Table 1, we report differences between households that were exposed to a sub-county deliberation baraza and households that were not. For this treatment, we can not reject balance on any of the variables. In the last column, we report differences in outcomes between households that were exposed to a district level baraza and households that were exposed to a sub-county level baraza that combined both information and deliberation components. We see that household heads in the treatment group are slightly older than in the control group. Furthermore, the share of households that report that there is a Village Health Team in their village is also slightly higher in the treatment group. In both cases, the difference are significant at a 10 percent level. Overall, out of 30 comparisons, we find that one difference is significant a the 5 percent level and three are significant at the 10 percent level, which is what one would expect to find due to chance alone. As such, we conclude that the initial randomization was successful.

5 Implementation challenges and subsequent design adaptations

One of the main challenges was a slow roll-out of the Baraza intervention by the implementing partner. At the start of 2018, and almost 2 and a half year after

Table 1: Orthogonality tests

	mean	sc baraza	information	deliberation	level
Household size	6.324	-0.284	0.304*	-0.003	0.246
	(2.825)	(0.190)	(0.133)	(0.125)	(0.248)
Age of the household head (years)	46.500	-0.474	0.464	0.725	1.432+
	(14.612)	(0.901)	(0.594)	(0.714)	(0.801)
Head of household is woman (1=yes)	0.191	0.023	-0.014	0.004	-0.013
	(0.393)	(0.021)	(0.013)	(0.015)	(0.016)
Head finished primary education (1=yes)	0.213	0.016	-0.020	-0.003	-0.026
	(0.410)	(0.027)	(0.020)	(0.020)	(0.027)
Thatched grass roof (1=yes)	0.298	0.021	0.009	-0.032	0.011
	(0.457)	(0.034)	(0.025)	(0.023)	(0.036)
Traditional mud wall (1=yes)	0.424	0.007	-0.025	0.038	-0.034
	(0.494)	(0.058)	(0.040)	(0.039)	(0.104)
Distance to nearest all weather road (km)	0.906	-0.088	0.106	0.147	-0.192
	(0.915)	(0.145)	(0.095)	(0.092)	(0.138)
Access to extension (1=yes)	0.108	-0.008	0.004	0.007	0.009
	(0.310)	(0.018)	(0.012)	(0.014)	(0.017)
Village Health Team in village (1=yes)	0.854	-0.031	0.006	0.025	0.070+
	(0.353)	(0.039)	(0.026)	(0.026)	(0.036)
Number of children in public schools	2.478	-0.161	0.165+	0.038	0.139
	(2.074)	(0.128)	(0.091)	(0.089)	(0.155)
Number of observations	12545	10241	10241	10241	4949

Note: First column reports sample means (and standard deviations below); Column 2 reports effect (and standard errors below) of the subcounty level baraza intervention; Column 3 reports the effect (and standard errors below) of the information component of the baraza intervention; Column 4 reports the effect (and standard errors below) of the administrative placement of the baraza intervention. **, * and + denotes significance at the 1, 5 and 10 percent levels.

baseline data was collected, only about 25 percent of the planned interventions had happened, and we needed to balance the pro and cons of pre-maturely ending the study or waiting for the implementing partner to implement all remaining barazas. At that time, we developed various scenarios, each with an adapted research design. After an additional six months, with still only 56 out of the 155 Barazas implemented, it appeared that the scenario whereby endline data would be collected before all subcounties were treated and potential selection bias introduced by a partial roll-out would be tested and accounted for during the analysis stage by relying on a matched difference-in-difference estimator.

However, end-line data collection after partial roll-out may introduce selection bias. It may be that from the randomly assigned sub-counties, particular sub-counties were selected to be treated first and others postponed. For instance, the implementing partner may have started with sub-counties that are close to the capital due to logistical reasons. In this note, we outline three strategies we will follow to address the issue of potential selection bias.

Third, during the analysis stage, the focus will now be on quasi-experimental evaluation designs – a matched difference-in-difference – in order to improve power and further reduce any remaining potential bias due to the partial roll-out. In the proposal, the focus was on average treatment effects, and for the power calculations, no co-variates were included. We thus expect substantial power improvements from controlling for pre-treatment differences in outcomes.

5.1 Balance between planned-to-treat-but-not-treated sub-counties, and control sub-counties

First, we can investigate if selection bias was introduced by comparing outcomes in control sub-counties to outcomes in sub-counties that were allocated to receive treatment but did not end up receiving treatment. The idea is that if the roll out was random, sub-counties that were allocated randomly to a particular treatment at the design stage but did not end up receiving treatment can be interchanged with sub-counties that were randomly selected at design stage to function as control sub-counties. Finding no significant differences in outcomes between these two groups would support the hypothesis that the partial roll-out did not introduce selection bias. We will already check this using pre-treatment characteristics collected during the baseline survey. During end-line data collections, we will also swap some of the control sub-counties with sub-counties that were supposed to be treated but did not end up receiving the treatment and test for selection bias using end-line data. If the incomplete roll-out introduced selection bias, comparing these two groups may also be informative to assess the direction and magnitude of the bias.

Table 2 presents the original balance table (Table 1), but after dropping sub-counties that were treated. Thus, instead of comparing pre-treatment characteristics between treatment sub-counties and control sub-counties, the table compares sub-counties that were allocated to a particular treatment (but did not end up receiving the treatment) to the control. The table seems to suggest

that the roll-out did not introduce imbalance, at least as judged by the pre-treatment characteristics that were in the original balance table. We find that, out of 20 comparisons, we reject the null hypothesis of no difference at a 5 percent significance level twice and at 10 percent once. While one may argue that, based on chance alone, one would expect to find only one significant difference at 5 percent, it should be noted that the two significant outcomes outcomes are likely to be positively correlated.

Furthermore, this can again be tested at end-line. There, we would expect to find similar post-treatment outcomes in the subset of sub-counties where a baraza was planned and not implemented, as well as in the subset of sub-counties where a Baraza was not planned (control Barazas).

5.2 Selection of control subcounties to be included in end-line survey

Second, as only part of the intervention was implemented, it will not be cost effective to collect end-line data on all sub-counties that did not receive a treatment (either because they were allocated to the control or because they ended up not being treated). Indeed, statistical power is likely to be highest for an equal number of treated and control sub-counties, and while adding more control sub-counties will increase precision, the gains are unlikely to outweigh the cost. This raises the question: from the potential control sub-counties (either those that were allocated to the control or because they ended up not being treated), which control sub-counties should be included in the data collection? One reasonable suggestion would be to pick them randomly. However, if the roll-out was not random, such a strategy may lead to a biased estimate of the causal impact of the intervention. For example, it may be that the implementer prioritized sub-counties that were closer to the capital. Randomly selecting control sub-counties may mean that sub-counties closer to the capital are relatively under-represented and sub-counties that are further away may be relatively over-represented in the control group. A better strategy may be to match, ex ante, each treated sub-county to a control sub-county that is similar in a range of observable pre-treatment characteristics that the planner had access to when rolling out the intervention and are likely to affect his or her decision. For instance, on the basis of the GPS coordinates of a treated sub-county, a control sub-county that is relatively close to the treated sub-county can be selected from the different candidate control sub-counties. It should be noted, though, that the reduction in bias of the treatment effect is likely to come at some cost in terms of external validity.

Figure 1 summarizes the factorial design that underlies the assessment of the relative effectiveness of the information and deliberation components of subcounty level barazas. One of the main advantages of factorial designs (as opposed to parallel designs) is the fact that, to test main effects, all observations can be used. For instance, to test the impact of an information Baraza, we can compare outcomes of households in sub-counties that received the information treatment (either only the information treatment or the information + deliber-

Table 2: Balance between planned but not treated subcounties and planned controls

	mean	sc baraza	information	deliberation
Household size	6.324	-0.315	0.388*	0.023
	(2.825)	(0.215)	(0.170)	(0.140)
Age of the household head (years)	46.500	-0.658	0.691	0.567
	(14.612)	(0.933)	(0.664)	(0.808)
Head of household is woman (1=yes)	0.191	0.020	-0.019	-0.003
	(0.393)	(0.023)	(0.016)	(0.017)
Head finished primary education (1=yes)	0.213	0.016	-0.007	-0.003
	(0.410)	(0.028)	(0.027)	(0.022)
Thatched grass roof (1=yes)	0.298	0.017	0.000	-0.036
	(0.457)	(0.036)	(0.024)	(0.027)
Traditional mud wall (1=yes)	0.424	0.001	-0.058	0.044
	(0.494)	(0.064)	(0.047)	(0.044)
Distance to nearest all weather road (km)	0.906	0.037	0.009	0.188+
	(0.915)	(0.165)	(0.100)	(0.110)
Access to extension (1=yes)	0.108	-0.002	0.008	0.007
	(0.310)	(0.019)	(0.016)	(0.015)
Village Health Team in village (1=yes)	0.854	-0.032	-0.010	-0.014
	(0.353)	(0.043)	(0.028)	(0.028)
Number of children in public schools	2.478	-0.176	0.249*	0.076
	(2.074)	(0.143)	(0.115)	(0.100)
Number of observations	12545	9241	7792	8341

Note: First column reports sample means (and standard deviations below); Column 2 reports differences (and standard errors below) of the information component of the baraza intervention; Column 3 reports differences (and standard errors below) of the information component of the baraza intervention; Column 4 reports differences (and standard errors below) of the administrative placement of the baraza intervention. **, * and + denotes significance at the 1, 5 and 10 percent levels.

	Control	Information
Control	Planned: 51 Included: 20	Planned: 51 Treated: 29
Deliberation	Planned: 51 Treated: 18	Planned: 53 Treated: 20

Figure 1: Factorial design

ation treatment) to outcomes of households that did not receive the information treatment (either because they received no treatment at all or because they only got the deliberation treatment). If the intervention had been implemented as planned, we would thus have had 104 information sub-counties that could be compared to 102 control sub-counties (and as 50 households were interviewed in each sub-county, we would have 5,200 treated households and 5,100 control households).

However, as mentioned above, the incomplete rollout resulted in the fact that only 67 of a total of 155 sub-counties that would have received any treatment were actually treated. Referring to Figure 1, we see that to test the impact of the information Baraza, 49 sub-counties that were treated can be used. This means that a total of 157 sub-counties that did not receive the information treatment can be used as control sub-counties. However, optimal power is obtained in designs where the number of treated units is about equal to the number of control units, so from a cost-efficiency perspective, it does not make sense to collect data on all of the 157 potential control sub-counties.

We will thus collect information on only a subset of the potential control sub-counties. In particular, to be able to test the impact of the information treatments, we will collect information on 54 sub-counties. As we want to formally test if the partial roll-out introduced selection bias by comparing planned control sub-counties to sub-counties that were not treated (see above), we will make sure we select half of these from the first column in table 2, and half from the second column1. To test the impact of the deliberation treatment, we need 38 control households. Also here, we will make sure half are from the planned controls (first row in table 2) and half from sub-counties that were supposed to be treated, but were not (second row). Finally, as we also plan to directly test for the effect of a combined information+deliberation treatment, we will need at least 20 pure control sub-counties. Also here, we will make sure half are selected from the upper left cell in table 2 and half from the sub-counties that were assigned to the treatment in the lower right cell of table 2 but did not get the treatment. Note that often, the same sub-county can be used to test different hypotheses. For instance, the 10 sub-counties in upper left cell needed to test if the deliberation intervention was effective can be taken from the 14 sub-counties that are needed in that cell to test the impact of the information treatment. We thus simply take the higher number in each cell, which is 14

sub-counties. To allow for attrition, we will select 16 control sub-counties in each treatment cell.

The question then becomes which control sub-counties to select for data collection from the potential control sub-counties in each cell. One way to proceed would be to simply pick at random. However, this may not be an optimal strategy if the roll-out did not happen randomly. For instance, it may be that sub-counties on the list that had connections to government (e.g. by having an MP or minister coming from that sub-county), were given priority in the roll-out. Simply picking comparison sub-counties randomly may result in a situation where only sub-counties without connections have been chosen, and it becomes difficult to separate a potential Baraza effect from the effect of having connections with government. It would be better to compare a treated and a non-treated sub-county with similar connections to the government. Therefore, we decide to use a range of sub-county characteristics that were likely to be known to the planner and may have affected how the intervention was rolled out to match each treated sub-county to a control sub-county that was similar in terms of these characteristics.

More in particular, we match on the following characteristics that were obtained at baseline from a survey of village chairs and chief administrative officers (CAO) of each sub-county: GPS coordinates of the sub-county, road infrastructure within the sub-county (km tarmac road and km all-weather (gravel) road), share of households with electricity, share of households with an iron roof or tiles, number of health centers in the sub-county, female primary school dropout rate, number of Universal Primary Education (UPE) schools in the sub-county, percent of farmers that use improved seed, and political connections of the sub-county (defined by having a minister or member of parliament coming from the sub-county). These characteristics are used in a probit regression to predict the likelihood that a sub-county was treated. For each treated sub-county, we then match a potential control sub-county with a likelihood of being treated that is similar to that of the treated sub-county. Using this method, the following (district - sub-counties) were selected in each treatment cell:

Note also that from the matched control sub-counties in the top right and bottom left cells, some may have already been included in the data collection if they were treated with the orthogonal factor. As a result, we only need to collect data on the 42 subcounties that did not received the intended treatment. Adding to this all these treated sub-counties, we get a total of 114 sub-counties we will collect data on, corresponding to about 5,700 households.

We follow the following procedure:

match 10 untreated sc from the 0 0 group to the 1 1 group

match 10 untreated sc from the 1 1 group to the 1 1 group

Next, we look at the information treatment. In this treatment, 54 have been treated, either as information alone or as part of the inf+deliberation treatment. This means we also need 54 controls. We already have selected 20 pure controls in the previous step which we can use. Furthermore, 18 pure deliberation treatments can be used as controls for the information treatment as well. This means we need an additional 16 controls. As we want to investigate

balance between control and planned but not treated controls, we select these 16 controls from the non treated subcounties that were planned to receive the information treatment.

Finally, we look at the deliberation treatment. In this treatment, 38 have been treated, either as deliberation alone or as part of the infd+delib treatment, so we also need 38 controls. We already have the 20 pure controls and and addtional 16 controls from the previous. So we need an additional 2 controls. As we want to investigate balance between control and planned but not treated controls, we select these 2 controls from the non treated subcounties that were planned to receive the deliberation treatment; we can.

Again, we look at baseline balance for this updated sample (Table 3). Interestingly, the imbalance that was found in Table 1 for the information treatment on household size and the number of children in school has disappear: apparently, the subcounties responsible for this imbalance ended up not being treated (Table 2).

5.3 Analysis: matched difference-in-difference

The revised approach for the impact evaluation of the Baraza initiative analysis entails an enhanced variant of difference-in-differences treatment effect estimations of our outcome indicators, drawing on both the baseline survey conducted in 2015, and the end-line survey data to be collected in 2019. The enhancement to the difference-in-differences (DID) estimation approach consists of applying kernel propensity score matching to the DID.

In our eventual panel data-set, each observation is recorded both before and after initiation of the Baraza project, and this is captured by a binary period variable, with 0 for the time period of the baseline period, before the start of Baraza implementation, and 1 for the end-line survey period, after Barazas have been conducted. The other key indicator is the treatment variable, which takes on the value of 0 for all households in the control sub-counties, and 1 for those in the treatment sub-counties (with treatments defined either as the main effects (information or deliberation) and the combined effect (information + deliberation)).

We furthermore specify several pre-treatment covariates of the model. These variables are used to estimate (using a probit regression) the propensity score for each observational unit, i.e. at the household level. This score is a single-index variable that reflects the likelihood of the unit being selected for treatment. The true treated and control units are then matched based on having similar propensity scores. Such a matching procedure is a statistical approach that can lower the extent to which there are differences between the control and treatment observations—in our case, at baseline. The type of matching method we employ, propensity score matching, is among the most commonly used in this class of quasi-experimental techniques. In our results, we present the inference from the first stage estimation of the propensity score. Weights are derived from this matching process and incorporated.

There are multiple ways to match treated and control units based on the

Table 3: Orthogonality tests for final sample

	mean	information	deliberation	level
Household size	6.407	0.019	0.065	-0.300^{+}
	(2.856)	(0.243)	(0.153)	(0.166)
Age of the household head (years)	46.975	0.275	-0.203	0.620
	(14.545)	(1.398)	(0.732)	(1.042)
Head of household is woman (1=yes)	0.191	0.009	-0.006	0.022
	(0.393)	(0.031)	(0.018)	(0.024)
Head finished primary education (1=yes)	0.208	0.014	-0.016	0.015
	(0.406)	(0.042)	(0.025)	(0.035)
Thatched grass roof (1=yes)	0.260	-0.024	0.044	-0.006
	(0.439)	(0.044)	(0.030)	(0.022)
Traditional mud wall (1=yes)	0.442	-0.027	0.031	0.063
	(0.497)	(0.085)	(0.053)	(0.058)
Distance to nearest all weather road (km)	0.905	-0.205	0.027	-0.103
	(0.910)	(0.198)	(0.140)	(0.135)
Access to extension (1=yes)	0.105	-0.007	0.000	0.013
	(0.307)	(0.025)	(0.012)	(0.020)
Village Health Team in village (1=yes)	0.863	-0.107^{+}	0.020	0.091*
	(0.344)	(0.059)	(0.036)	(0.039)
Number of children in public schools	2.506	0.091	0.001	-0.188^{+}
	(2.073)	(0.163)	(0.097)	(0.110)
Number of observations	12545	5348	5348	5348

Note: First column reports sample means (and standard deviations below); Column 2 reports differences (and standard errors below) of the information component of the baraza intervention; Column 3 reports differences (and standard errors below) of the information component of the baraza intervention; Column 4 reports differences (and standard errors below) of the administrative placement of the baraza intervention. **, * and + denotes significance at the 1, 5 and 10 percent levels.

similarity of their propensity scores. Since no two propensity scores are identical, we use nearest neighour matching. Specifically, for any given treated unit, the weight applied to each control unit is inversely proportional to the difference between the propensity scores of these two units. In other words, the greater the gap between the two propensity scores, the smaller the weight applied to that control unit.

We will be able to demonstrate the extent to which several variables correlate with the likelihood of being selected into treatment. For doing so, we will consider core household demographics, household head education, and household farm size—factors that can be expected to affect treatment. Even if in our case treatment is at the sub-county level, the purposive selection of treatment sub-counties can be expected to correlate with average household characteristics of the sub-counties.

We will also perform a balancing test analysis, and examine the results of the balancing t-test of the difference in the means of the weighted covariates between the control and treated groups at the time of the baseline. The balancing properties will be tested anew for each separate regression. By presenting this side by side with the unmatched balance, we will be able to examine whether there is an improvement in the baseline balance after matching, to show the extent to which selection bias is reduced by matching.

The analysis will be carried out for all observations, as well as on only the common support of the propensity score. The latter means that the balancing test is performed only on those observations whose propensity scores lie within the range of propensity scores of the matched treated and control units. Doing so can improve the quality of the matches used to estimate the average treatment effect, even if this means that the number of observations used for this estimation is somewhat reduced. If our matching is successful, we would find that after matching, there is no statistically significant difference between the means of the covariates at baseline, ideally in the balancing analysis drawing on all data as well as only on those observations whose propensity scores are within the common support range.

The average treatment effect will then be estimated using DID, and through regression in a linear probability model—as mentioned, both on the full dataset as well on only the common support of the propensity score.

6 Updated power calculations

The original power calculations assumed full roll-out of the intervention. The partial roll-out necessitates an update to these power calculations to obtain a new set of minimal detectable effects (MDEs) associated with the sample that will be collected. Below, we use baseline data to simulate MDEs for a selection of the outcomes we will use to judge effeciveness of the intervention (and are described in detail in section 7). We use a standard significance level of 0.05 (double sided).

Figure 2 plots MDEs against power for the first outcome variable that will

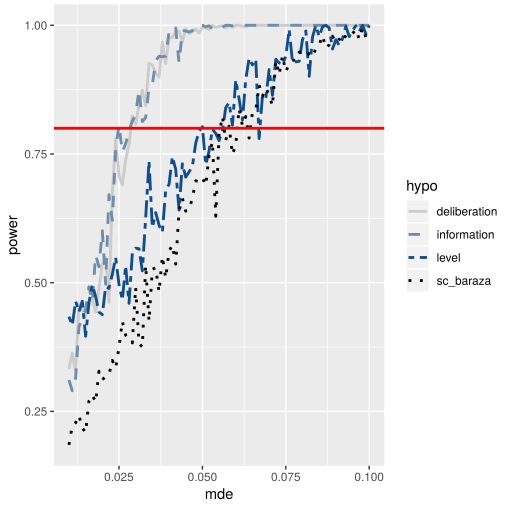

Figure 2: Power curves for access to extension

be used to assess changes impact on public service delivery in the agricultural sector (extension at home, measured as the percentage of households in our sample who report that they were visited by an expert). On average, about 11 percent of households in our sample report that they were visited by an extension officer in the last year. The gray solid line shows the power curve associated to the deliberation treatment, comparing the 1,900 households that received the information treatment to the 3,450 households that did not receive a deliberation focused baraza. The blue dashed line shows power for different MDEs for the information component of the baraza intervention. Here, we compare the 2,450 households that live in subcounties that received an information baraza to the 2,900 households that did not receive a subcounty information baraza. Finally, the dotted dark blue line compares effectiveness of barazas conducted at different levels, with the MDE defined as the difference in outcome between 1,000 households that received the combined information and deliberation subcounty level baraza and 2,000 households that were exposed to a district level baraza. MDEs are estimated using a simple ANCOVA model that controls for the outcome at baseline. We modeled an autocorrelation of 10 percent.

Not surprisingly, we have most power for testing the information treatment. We see that the power curve hits the 80 percent threshold a first time at an MDE of about 2.5. The deliberation experiment is similarly powered, and at 80 percent we can expect to identify effects of 3 percent of more. Due to the smaller sample size and more aggregated treatment, directly comparing subcounty level barazas to district level barazas seems harder. Here, the difference needs to be at least 5 percent[4].

[4]Unless effects from district level barazas and sub-county level barazas go in opposite directions, it seems unlikely that we will be able to detect a difference in a direct comparison. Even

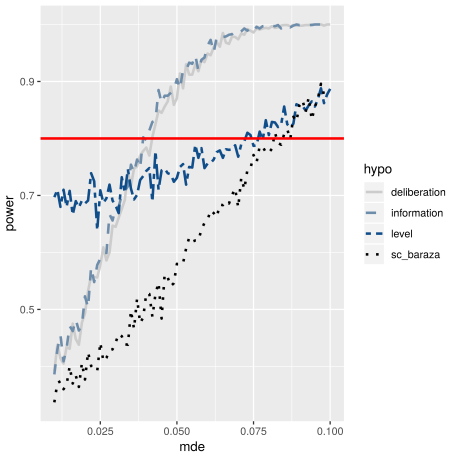

Figure 3: Power curves for distance to water source

In figure 3, we plot MDEs for an infrastructure related outcome: distance (in km) to the primary water source during the dry season. We find that for the information treatment and the deliberation treatment, we can detect an 4 percent difference at the standard 80 percent power level. As the average household lives about 900 meters from this means we can identify effects in excess of 36 meters. Also here, the MDE is higher when directly comparing the effect of district level barazas to subcounty level barazas. This MDE corresponds to about 70 meters for the average household in our sample.

In Appendix XXX, we run similar analysis for all the variables that we will use to judge impact of the baraza intervention.

7 Results

In this section, we provide results for the three main hypotheses outlined in Section 3. For now, the tables and graphs are generated by running the code on a simulated endline data set[5]. Once endline data is collected, we simply replace the simulated endline dataset with the actual endline. Preparing and pre-registering such a "mock report" should reduce the likelihood that results are driven by specification search.

For continuous variables, 5 percent trimmed values will be use (2.5 percent trimming at each side of the distribution). Inverse hyperbolic sine transforma-

though a direct comparison is most interesting and most relevant from a policy perspective, we have more power for other comparisons. For instance, we can also compare outcomes of the 2,000 households that were exposed to a district level baraza to the 2,000 households that did not receive a baraza. Alternatively, we can compare outcomes of the 2,000 households that were exposed to a district level baraza to the 3,350 households that received any type of sub-county level baraza (information only, deliberation only, or the combined).

[5]The simulated endline was just a random draw from the baseline data

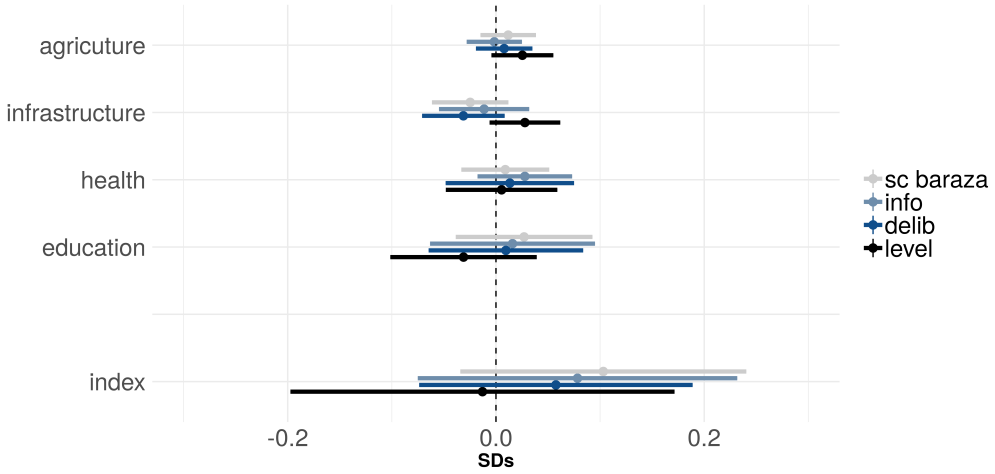

Figure 4: Summary of baraza impact

tions will be used if skewness exceeds 1.96. Trimming will always be done on end results. For instance, if the outcome is yield at the plot level, then production will first be divided by plot area, after which logarithms are taken and the end result is trimmed. Outcomes for which 95 percent of observations have the same value within the relevant sample will be omitted from the analysis and will not be included in any indicators or hypothesis tests.

Figure 4 provides a summary of there results from the baraza impact evaluation. It shows the impact of the three key attributes of the baraza intervention–the information aspect, the deliberation component, and the administrative level at which the baraza took place–on four sectors we consider. The graphs are based on indices that are composed of individual outcomes in each sector. We also combine the four indices into one overall index that assesses the impact on public service delivery in general.

7.1 ANCOVA

In this section, we present results for a simple treatment-control comparison, implemented using an OLS regression that also controls for the region (as this was used for stratification) and the baseline outcome. In particular, we estimate the following regression models for each outcome variable.

7.1.1 Agriculture:

The following outcomes will be considered in assessing the effectiveness of baraza to change service delivery in agriculture:

- extension at home: measured as the percentage of households in our sample who report that they were visited by an expert (e.g. crop or livestock extension agent, or community based facilitator or another experienced farmer) at the home in the last 12 months (baraza.B2).
- extension visits: measured as the percentage of households in our sample who report that someone in the household visited an extension office or a meeting/training organized by an extension officer or visited a demonstration site/ model farmer (baraza.B3 or baraza.B3.3).
- NAADS in village: measured as the percentage of households in our sample that reports that presence of a farmer association/group supported by NAADS or Operation Wealth Creation (baraza.B4.1).
- input use: measured as the proportion of households in our sample that report using inorganic fertilizers (DAP, Urea, NPK, Foliar, TSP, SSP, MOP) or improved seed in the last 12 months (baraza.B1 or baraza.B1.5).
- marketing help village committee: measured as the proportion of households in our sample that report they received help in marketing their produce from the village procurement committee/village farmers forum in the last 12 months. (baraza.B5.2)
- marketing help cooperative: measured as the proportion of households in our sample that report they received help in marketing their produce from the Cooperative/Association in the last 12 months. (baraza.B5.3)

```
## [1] 1
## [1] 2
## [1] 3
## [1] 4
## [1] 5
## [1] 6
## [1] 7
## [1] 8
## [1] 9
## [1] 10
## [1] 11
## [1] 12
## [1] 13
## [1] 14
## [1] 15
## [1] 16
## [1] 17
## [1] 18
## [1] 19
## [1] 20
```

```
## [1] 21
## [1] 22
## [1] 23
## [1] 24
## [1] 25
## [1] 26
## [1] 27
## [1] 28
## [1] 29
## [1] 30
```

7.1.2 Infrastructure

This study focus was primarily on water and road infrastructure. A first outcome we consider is whether the household uses unprotected water source during dry season (yes/no). This is measured as the share of households that report that the main source of drinking water during the dry season is rain water, surface water, water obtained from a tube well or borehole, an unprotected dug well or and unprotected spring. (baraza.C1). We find that overall, 70 percent of households rely on unsafe water during the dry season.

- distance to water source: measured as the distance in km of the primary source during the dry season (baraza.C1.2)
- waiting time: measured as the average time in minutes one has to wait to collect water at the primary source during the dry season? (baraza.C1.3)
- ##[11,] "baraza.C2.3" "c10" "Is there a Water User Committee in this village?" #
- #[12,] "baraza.A6" "a6" "Distance to nearest all weather road (km)"

Table 5 provides details of the impact of the baraza intervention on infrastructure.

7.1.3 Health

- Health: ##[14,] "baraza.D2" "pub_health_access" "Seek treatment for fever in public health facility"
- ##[15,] "baraza.D2.4" "maternal_health_access" "Go to public health facility to give birth"
- ##[16,] "baraza.D3" "d31" "Is there a VHT in village"
- ##[17,] "baraza.D4.2" "d43" "Distance to nearest govt health facility"
- ##[18,] "baraza.D1" "d11" "Were any household members unable to work or go to school due to an illness in the past one year?"

Table 4: Treatment-control differences (ANCOVA) - Agriculture

	mean	sc baraza	information	deliberation	level
Was visited by extension officer at home (yes/no)	0.112	0.014	0.000	0.005	-0.007
	(0.316)	(0.017)	(0.010)	(0.008)	(0.010)
Visited training or demonstration site (yes/no)	0.080	0.002	-0.008	0.000	-0.008
	(0.271)	(0.014)	(0.010)	(0.013)	(0.009)
NAADS or OWC in village (yes/no)	0.179	-0.029*	0.019	-0.006	0.012
	(0.384)	(0.012)	(0.012)	(0.016)	(0.013)
Uses modern inputs (improved seed or fertilizer) (yes/no)	0.339	-0.004	0.002	-0.003	0.001
	(0.473)	(0.013)	(0.013)	(0.015)	(0.016)
Support in marketing from village procurement committe (yes/no)	0.439	-0.027+	0.004	0.008	0.015
	(0.496)	(0.015)	(0.016)	(0.018)	(0.022)
Support in marketing from cooperative (yes/no)	0.269	0.010	0.016	-0.008	0.017
	(0.444)	(0.014)	(0.014)	(0.015)	(0.015)
Agriculture Index	0.000	-0.009	0.012	-0.002	0.008
	(0.409)	(0.014)	(0.013)	(0.013)	(0.013)

Note: First column reports sample means (and standard deviations below); Column 2 reports differences (and standard errors below)

Table 5: Treatment-control differences (ANCOVA) - Infrastructure

	mean	information	deliberation	level
Household uses unprotected water source during dry season (yes/no)	0.696	0.013	0.020	-0.012
	(0.460)	(0.014)	(0.018)	(0.019)
Distance to water source	0.708	-0.006	-0.009	-0.006
	(0.529)	(0.014)	(0.021)	(0.020)
Average waiting time at source (min)	2.678	-0.012	0.057	-0.021
	(2.179)	(0.017)	(0.073)	(0.092)
Is there a Water User Committee in this village? (yes/no)	0.528	-0.011	-0.023	-0.028
	(0.499)	(0.015)	(0.016)	(0.020)
Distance to nearest all weather road (km)	0.912	-0.007	0.004	-0.037
	(0.918)	(0.013)	(0.028)	(0.041)
Infrastructure Index	0.000	-0.018	-0.025	-0.011
	(0.442)	(0.016)	(0.018)	(0.021)

Note: First column reports sample means (and standard deviations below); Column 2 reports differences (and standard errors below)

- ##[19,] "baraza.D4.6" "wait_time" "How long did you have to wait before being attended (in min)"
- #20 "baraza.D6" "d61" "Have you visited a traditional health practitioner in the last year?"

7.1.4 Education

- ##[22,] "n_children" "base_n_children" "Number of children in UPS or USE"
- ##23"baraza.E1" e5upe "Distance to public school
- ##24"baraza.E1.4" baseline$e12upe "Complete boundary fence
- ##25"baraza.E1.6" baseline$e14upe "Has water facility
- ##26"baraza.E1.10 baseline$e22upe "Has SMC
- ##27"baraza.E1.13 baseline$e32upe "Informed about SMC
- ##28"baraza.E1.18 e45upe "Inspectors visited schools

7.2 matched difference-in-difference

We also look at how impact changes over time. A graphical approach uses non-parametric smoothing to look at the average impact depending on the time between treatment assignment and endline data collection.

8 Threat to validity

The main threat to validity stems from the risk that incomplete roll-out introduced selection bias into the study. While the results in Table 2 are reassuring, it should be noted that pre-treatment characteristics were collected some time ago and results may be different if more recent data is used and/or if selection happened on characteristics that change over time. Therefore, comparing characteristics between control sub-counties and sub-counties that were allocated to a treatment cell but ended up not being treated will be an important strategy to test for selection bias that may have been introduced by the incomplete roll-out of the treatment. In particular, instead of simply collecting information from the (planned) control sub-counties, we will also collect information from sub-counties that were supposed to receive a treatment but did not get one. If the incomplete roll-out introduced selection bias, this information may also be useful to assess the direction and magnitude of the bias.

Table 9 thus compares endline outcomes between households that were planned to receive a particular treatment but did not end up receiving the treatment to

Table 6: Treatment-control differences (ANCOVA) - Health

	mean	information	deliberation	level
Seek treatment for fever in public health facility (1=yes)	0.737	0.022	0.007	-0.009
	(0.440)	(0.016)	(0.015)	(0.018)
Go to public health facility to give birth (1=yes)	0.841	0.016	-0.004	-0.004
	(0.366)	(0.014)	(0.011)	(0.013)
Is there a VHT in village? (1=yes)	0.755	0.011	0.033*	0.027
	(0.430)	(0.019)	(0.016)	(0.017)
Distance to nearest govt health facility (km)	1.995	-0.029*	-0.014	-0.023
	(0.743)	(0.014)	(0.025)	(0.032)
Were days work/school missed due to illness? (1=yes)	0.792	0.004	-0.008	0.012
	(0.406)	(0.014)	(0.014)	(0.016)
Waiting time before being attended (min)	4.599	-0.003	-0.032	-0.095
	(1.341)	(0.020)	(0.067)	(0.080)
Has visited traditional health practitioner? (1=yes)	0.091	-0.007	0.008	-0.004
	(0.288)	(0.013)	(0.009)	(0.010)
Health Index	0.000	-0.006	0.009	0.028
	(0.381)	(0.017)	(0.021)	(0.022)

Note: First column reports sample means (and standard deviations below); Column 2 reports differences (and standard errors below)

Table 7: Treatment-control differences (ANCOVA) - Education

	mean	information	deliberation	level
Number of children in UPS or USE	2.481	-0.014	-0.009	0.013
	(2.065)	(0.012)	(0.067)	(0.089)
Distance to public school (km)	1.411	-0.002	-0.051	-0.050
	(0.757)	(0.026)	(0.048)	(0.054)
Has complete boundary fence (1=yes)	0.313	0.002	0.010	0.002
	(0.464)	(0.015)	(0.019)	(0.021)
Has water facility (1=yes)	0.681	-0.008	-0.030	0.004
	(0.466)	(0.019)	(0.022)	(0.027)
Has SMC (1=yes)	0.767	-0.005	-0.002	0.004
	(0.422)	(0.016)	(0.017)	(0.021)
Is informed about SMC (1=yes)	0.351	0.041*	-0.029	-0.026
	(0.477)	(0.020)	(0.023)	(0.026)
Inspectors visited schools (1=yes)	0.507	0.018	-0.043*	-0.013
	(0.500)	(0.022)	(0.022)	(0.030)
Education Index	0.000	0.018	0.027	0.016
	(0.376)	(0.033)	(0.033)	(0.038)

Note: First column reports sample means (and standard deviations below); Column 2 reports differences (and standard errors below)

Table 8: Matched difference in difference - Agriculture

	mean	information	deliberation	level
Was visited by extension officer at home (yes/no)	0.112	0.054	0.002	-0.017
	(0.316)	(0.050)	(0.020)	(0.024)
Visited training or demonstration site (yes/no)	0.080	0.036	0.009	0.000
	(0.271)	(0.044)	(0.019)	(0.023)
NAADS in village (yes/no)	0.179	0.016	0.034	-0.046
	(0.384)	(0.073)	(0.032)	(0.033)
Uses modern inputs (improved seed or fertilizer) (yes/no)	0.339	-0.065	0.010	-0.046
	(0.473)	(0.087)	(0.045)	(0.044)
Support in marketing from Village procurement committe (yes/no)	0.439	0.022	0.027	-0.017
	(0.496)	(0.100)	(0.043)	(0.043)
Support in marketing from cooperative (yes/no)	0.269	-0.051	0.016	0.003
	(0.444)	(0.087)	(0.051)	(0.049)
Agriculture Index	0.000	-0.032	0.041	-0.027
	(0.409)	(0.078)	(0.035)	(0.045)

Note: First column reports sample means (and standard deviations below); Column 2 reports differences (and standard errors below)

Table 9: Difference between planned but not treated subcounties and planned controls at endline

	information	deliberation
Agriculture index	0.007	-0.004
	(0.019)	(0.022)
Infrastructure index	0.030^{+}	-0.006
	(0.018)	(0.022)
Health index	0.021	0.003
	(0.018)	(0.023)
Education index	0.015	-0.039
	(0.019)	(0.024)
Public service delivery index	0.051	0.006
	(0.034)	(0.047)
Number of observations	2891	3440

Note: First column reports differences (and standard errors below) of the information component of the baraza intervention; Column 3 reports differences (and standard errors below) of the deliberation component of the baraza intervention; standard errors are clustered at the subcounty level; **, * and + denotes significance at the 1, 5 and 10 percent levels.

outcomes of households that were assigned to serve as a control for the particular treatment. In the table, we present results for the indices that are also used to summarize impact in Figure 4; full results with details for each outcome in the four sectors can be found in Appendix XXX.

9 Acknowledgment

This project received funding from the International Institute for Impact Evaluations (3ie) under grant number RFQ PW2.18.UG.PG. This research was partly funded by the CGIAR Research Program on Policies, Institutions, and Markets (PIM), led by the International Food Policy Research Institute (IFPRI) and carried out with support from the CGIAR Fund contributors (https://www.cgiar.org/funders/) under grant number PIM 203002.002.221. We would like to thank Alvina Erman, Jennifer Smart, Marc Charles Wanume, Wilberforce Walukano, Fiona Nattembo, Emmanuel Bizimungu, Leocardia Nabwire and people at the Office of the Prime Minister, Joseph Muserero in particular.

References

Anderson, M. L. 2008. "Multiple inference and gender differences in the effects of early intervention: A reevaluation of the Abecedarian, Perry Preschool,

and Early Training Projects." *Journal of the American statistical Association* 103 (484): 1481–1495.

Arkedis, J., J. Creighton, A. Dixit, A. Fung, S. Kosack, and D. Levy. 2019. "Can Transparency and Accountability Programs Improve Health? Experimental Evidence from Indonesia and Tanzania." .

Banerjee, A. V., E. Duflo, and R. Glennerster. 2008. "Putting a Band-Aid on a Corpse: Incentives for Nurses in the Indian Public Health Care System." *Journal of the European Economic Association* 6 (2-3): 487–500.

Björkman, M. and J. Svensson. 2009. "Power to the people: evidence from a randomized field experiment on community-based monitoring in Uganda." *The Quarterly Journal of Economics* 124 (2): 735–769.

Björkman Nyqvist, M., D. de Walque, and J. Svensson. 2017. "Experimental evidence on the long-run impact of community-based monitoring." *American Economic Journal: Applied Economics* 9 (1): 33–69.

Francis, P. and R. James. 2003. "Balancing rural poverty reduction and citizen participation: The contradictions of Uganda's decentralization program." *World Development* 31 (2): 325–337.

Fujiwara, T. and L. Wantchekon. 2013. "Can informed public deliberation overcome clientelism? Experimental evidence from Benin." *American Economic Journal: Applied Economics* 5 (4): 241–55.

Gilens, M. 2001. "Political Ignorance and Collective Policy Preferences." *The American Political Science Review* 95 (2): 379–396.

Goeree, J. K. and L. Yariv. 2011. "An Experimental Study of Collective Deliberation." *Econometrica* 79 (3): 893–921.

Golooba-Mutebi, F. 2005. "When popular participation won't improve service provision: primary health care in Uganda." *Development Policy Review* 23 (2): 165–182.

Grossman, G. and K. Michelitch. 2018. "Information dissemination, competitive pressure, and politician performance between elections: A field experiment in Uganda." *American Political Science Review* 112 (2): 280–301.

Gutmann, A. and D. F. Thompson. 1998. *Democracy and disagreement*. Harvard University Press.

Humphreys, M., R. S. De la Sierra, and P. Van der Windt. 2013. "Fishing, commitment, and communication: A proposal for comprehensive nonbinding research registration." *Political Analysis* 21 (1): 1–20.

Keefer, P. and S. Khemani. 2005. "Democracy, Public Expenditures, and the Poor: Understanding Political Incentives for Providing Public Services." *The World Bank Research Observer* 20 (1): 1–27.

Kosec, K. and L. Wantchekon. 2020. "Can information improve rural governance and service delivery?" *World Development* 125: 104376.

Olken, B. A. 2007. "Monitoring corruption: evidence from a field experiment in Indonesia." *Journal of political Economy* 115 (2): 200–249.

OPM. 2013. *Implementation of the Baraza Initiative. Progress Report.* Tech. rep.

Pandey, P., S. Goyal, and V. Sundararaman. 2009. "Community participation in public schools: impact of information campaigns in three Indian states." *Education Economics* 17 (3): 355–375.

Raffler, P., D. N. Posner, and D. Parkerson. 2018. "The weakness of bottom-up accountability: Experimental evidence from the Ugandan health sector." *Unpublished manuscript* .

Reinikka, R. and J. Svensson. 2004. "Local capture: evidence from a central government transfer program in Uganda." *The quarterly journal of economics* 119 (2): 679–705.

Van Campenhout, B., E. Bizimungu, J. Smart, and N. Kabunga. 2018. "Impact pathways of a participatory local governance initiative in Uganda: a qualitative exploration." *Development in Practice* 28 (8): 1046–1056.

10 Appendix One

11 Appendix Two

Figure 5: Ag MDEs

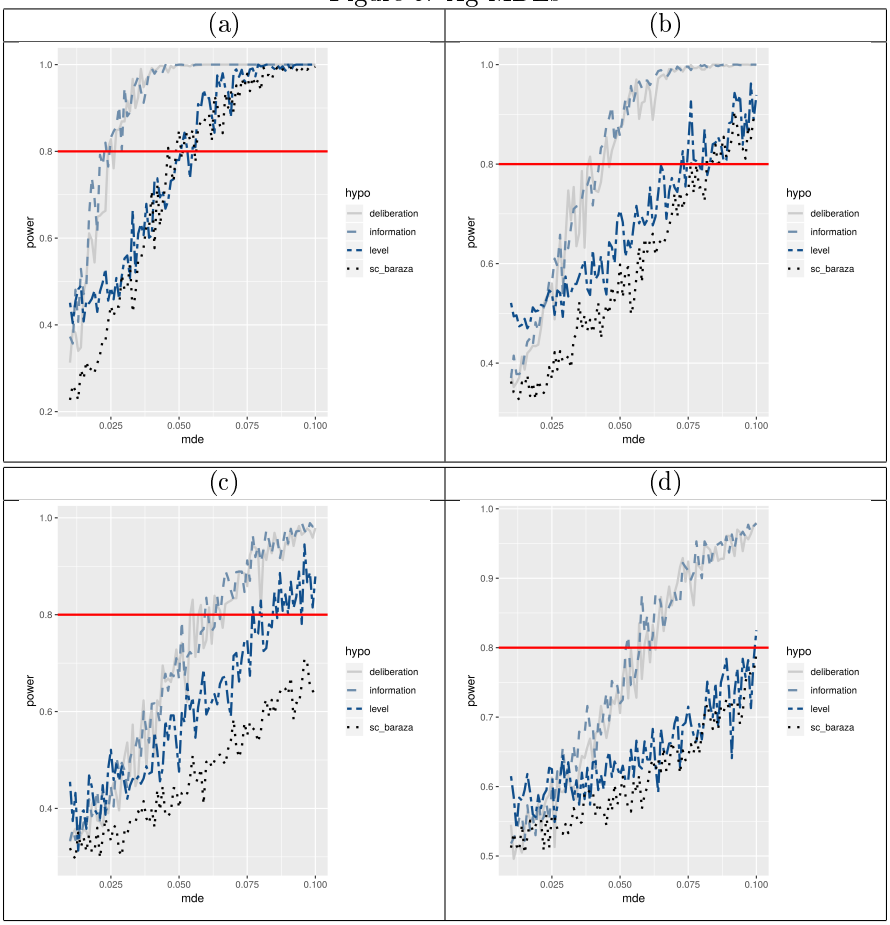

Table A.1: Difference between planned but not treated subcounties and planned control subcounties – Agriculture

	information	deliberation
Was visited by extension officer at home (yes/no)	0.003	0.011
	(0.012)	(0.016)
Visited training or demonstration site (yes/no)	0.006	-0.021
	(0.011)	(0.014)
NAADS or OWC in village (yes/no)	-0.010	-0.001
	(0.014)	(0.018)
Uses modern inputs (improved seed or fertilizer) (yes/no)	0.012	0.022
	(0.017)	(0.024)
Support in marketing from village procurement committe (yes/no)	0.009	0.023
	(0.020)	(0.023)
Support in marketing from cooperative (yes/no)	-0.003	-0.031
	(0.022)	(0.029)
Agriculture index	0.007	-0.004
	(0.019)	(0.022)
Number of observations	2891	3440

Note: First column reports differences (and standard errors below) of the information component of the baraza intervention; Column 3 reports differences (and standard errors below) of the deliberation component of the baraza intervention; standard errors are clustered at the subcounty level; **, * and + denotes significance at the 1, 5 and 10 percent levels.

Table A.2: Difference between planned but not treated subcounties and planned control subcounties – Infrastructure

	information	deliberation
Household uses unprotected water source during dry season (yes/no)	-0.038*	0.011
	(0.017)	(0.020)
Distance to water source	-0.004	-0.021
	(0.021)	(0.026)
Average waiting time at source (min)	0.121+	0.170
	(0.072)	(0.105)
Is there a Water User Committee in this village? (yes/no)	0.007	-0.020
	(0.016)	(0.022)
Distance to nearest all weather road (km)	0.047	0.005
	(0.038)	(0.045)
Infrastructure index	0.030+	-0.006
	(0.018)	(0.022)
Number of observations	2891	3440

Note: First column reports differences (and standard errors below) of the information component of the baraza intervention; Column 3 reports differences (and standard errors below) of the deliberation component of the baraza intervention; standard errors are clustered at the subcounty level; **, * and + denotes significance at the 1, 5 and 10 percent levels.

Table A.3: Difference between planned but not treated subcounties and planned control subcounties – Health

	information	deliberation
Seek treatment for fever in public health facility (1=yes)	0.011	-0.019
	(0.016)	(0.022)
Go to public health facility to give birth (1=yes)	0.032*	-0.017
	(0.013)	(0.015)
Is there a VHT in village? (1=yes)	-0.023	0.013
	(0.015)	(0.025)
Distance to nearest govt health facility (km)	-0.038	-0.033
	(0.028)	(0.037)
Were days work/school missed due to illness? (1=yes)	-0.010	-0.007
	(0.015)	(0.021)
Waiting time before being attended (min)	-0.038	0.049
	(0.056)	(0.089)
Has visited traditional health practitioner? (1=yes)	-0.013	0.023*
	(0.009)	(0.010)
Health index	0.021	0.003
	(0.018)	(0.023)
Number of observations	2891	3440

Note: First column reports differences (and standard errors below) of the information component of the baraza intervention; Column 3 reports differences (and standard errors below) of the deliberation component of the baraza intervention; standard errors are clustered at the subcounty level; **, * and + denotes significance at the 1, 5 and 10 percent levels.

Table A.4: Difference between planned but not treated subcounties and planned control subcounties – Education

	information	deliberation
Number of children in UPS or USE	0.058	-0.092
	(0.067)	(0.077)
Distance to public school (km)	-0.024	-0.017*
	(0.042)	(0.044)
Has complete boundary fence (1=yes)	-0.009	0.007
	(0.015)	(0.022)
Has water facility (1=yes)	-0.014	-0.035
	(0.021)	(0.026)
Has School Management Committee (1=yes)	0.009	0.029
	(0.018)	(0.020)
Is informed about School Management Committee (1=yes)	0.002	-0.006
	(0.020)	(0.021)
Inspectors visited schools (1=yes)	0.028	-0.055*
	(0.018)	(0.026)
Education index	0.015	-0.039
	(0.019)	(0.024)
Number of observations	2891	3440

Note: First column reports differences (and standard errors below) of the information component of the baraza intervention; Column 3 reports differences (and standard errors below) of the deliberation component of the baraza intervention; standard errors are clustered at the subcounty level; **, * and + denotes significance at the 1, 5 and 10 percent levels.