Social accountability and service delivery:

Experimental evidence from Uganda[1]

Nathan Fiala and Patrick Premand[2]

September 2019

Abstract

Corruption and mismanagement of public resources can affect the quality of government services and undermine growth. How can citizens in poor communities be empowered to demand better-quality public investments? We look at whether providing social accountability skills and information on project performance can lead to improvements in local development projects supported by a large national program. We find that offering communities training improves project output. The combination of training and information on project quality leads to significant and large improvements in household assets, while providing either social accountability training or project quality information by itself has no effects on household assets. We explore mechanisms and show that the impacts come in part from community members increasing their monitoring of local projects, making more complaints to local and central officials and increasing cooperation. We also find modest improvements in people's trust in the central government. The results suggest that government-led, large-scale social accountability programs can make development projects more effective and improve citizens' welfare.

JEL codes: D7, H4, O1

Keywords: Social accountability; community training; scorecards; corruption; service delivery

[1] This study was pre-registered under AEARCTR-0001115. We are very thankful to Suleiman Namara and Endashaw Tadesse, who led the design and supervision of the program at the World Bank; and James Penywii and Munira Ali, who managed it at the Ugandan Inspectorate of Government. We thank Filder Aryemo and Jillian Larsen for outstanding research and operational contributions; Iker Lekuona, Kalie Pierce, Simon Robertson, Areum Han and Mariajose Silva Vargas for excellent research assistance; the study participants for generously giving their time; as well as the field officers of Innovations for Poverty Action. Data collection was funded by a Vanguard Charitable Trust and the World Bank, including grants from the i2i and NTF Trust Funds. We are grateful for comments provided at various points during this study by Colin Andrews, Chris Blattman, Bénédicte de la Brière, Robert Chase, Deon Filmer, Vincenzo Di Maro, Christina Malmberg Calvo, Isabel Günther, Ezequiel Molina, Benjamin Olken, Obert Pimhidzai, Pia Raffler, Ritva Reinikka, Dena Ringold, Danila Serra and Lynne Sherburne-Benz, as well as audiences at Harvard University, Makerere University, GIGA, RWI, DIW Berlin, the University of Connecticut, ETH Zürich and the World Bank. All findings, interpretations, and conclusions in this paper are those of the authors and do not necessarily represent the views of the World Bank or the government of Uganda.

[2] Fiala: University of Connecticut, Makerere University and RWI–Leibniz Institute for Economic Research; nathan.fiala@uconn.edu. Premand: World Bank and ETH Zürich; ppremand@worldbank.org.

1 Introduction

Corruption and mismanagement of public resources can undermine development by generating costs for society. Those costs can range from an increase in bureaucratic hurdles to extract payments from citizens, to the creation of an unappealing economic environment for foreign investments, or a reduction of human capital stemming from low-quality delivery of health or education services (Bertrand et al., 2007; Woo, 2010; Reinikka and Svensson, 2004; Björkman and Svensson, 2009; Bold et al., 2017). Corruption and mismanagement can also increase inequality by affecting more severely those with less voice but greater need for public services (Olken, 2006; Hunt, 2007). Community and government officials may misuse or divert funds from local populations. When combined with collective action problems and lack of information and skills to address these issues, corruption could lead to significant problems in service delivery.

We explore whether and how citizens in poor communities can be empowered to demand better-quality public investments. We worked with the Government of Uganda to conduct an experiment with a large sample of communities across the north of the country. We test whether providing monitoring skills and encouraging the reporting of cases of mismanagement, as well as disseminating information on the absolute and relative performance of community projects, pushes citizens to demand and obtain better local development projects.

Communities were selected by the central government to receive a community-driven development program called the Second Northern Uganda Social Action Fund (NUSAF2). NUSAF2 comprised a wide range of project types, including building teachers' houses, providing livestock to households, putting up fencing, and enterprise development. The study took place in 940 communities that had already chosen a type of project and were awarded

NUSAF2 funding to implement it. We randomly selected 634 of these communities to receive a six-day training on how to monitor community projects, including how to identify and make complaints about corruption and mismanagement to implementing partners and local, sub-national, or national leaders. The trainings were managed by the Inspectorate of Government (IG), an independent arm of the government responsible for fighting corruption and implemented in partnership with local civil society organizations (CSOs).

We developed a normalized index of project quality obtained through physical assessments of the projects (similar to audits). These data were collected about one year after the start of the local NUSAF2 projects and were used to measure the immediate impacts of the training. To determine if training alone is enough or if it needs to be combined with information on how well communities perform, we then used the information collected from this assessment to create a scorecard that ranks the performance of the community projects relative to other community projects within a district. We randomly selected 283 communities to be given this information during a community meeting, which included a facilitated discussion about why communities did or did not perform well relative to others.

This produced a 2x2 design where communities received training, a scorecard, both training and a scorecard, or no intervention. This design allows us to test directly whether training communities on social accountability or simply providing information on relative project quality can improve the outputs from local development projects, or if a combination of both training and scorecard information is needed. As such, the focus is not on the effectiveness of the NUSAF2 project itself, but rather on whether incorporating social accountability training or scorecard information can improve the effectiveness of local development projects for communities and households.

Our experiment is embedded in a large-scale community-driven development program. The scale of the intervention is between five and 20 times larger than in similar research, covering 45 districts and 485 sub-counties throughout the northern half of Uganda, with more than 10,000 direct beneficiaries.[3] Given the large available sample, the design is well powered and allows for a minimum detectible effect size of less than 0.10 standard deviations for most outcomes.

We conducted individual surveys with community members six months after the initial assessment and scorecards were delivered to measure impacts at the household level. The sample includes over 6,900 individuals. Almost two-thirds of the projects provided livestock, making these projects more easily comparable to one another and more likely to have welfare implications for individual households. For these reasons, we focus our analysis of household impacts on communities that applied for livestock projects before the interventions were randomized, though we also present results from the full sample.

We find that the social accountability training led to a small increase in the project outputs by 0.119 standard deviations. From the follow-up household survey conducted six months later, we find that neither the training nor the project quality scorecard alone had any impact on household assets. However, the combination of the two led to large increases in household assets: households in communities that received both training and information scorecards have approximately 0.42 more head of cattle per household, or 19% more than the control group. This is equivalent to approximately $97 per household (or between $970 to $1,455 per community) worth of animals. These findings indicate that for rural Ugandans, who often have limited interactions with the government, providing training alone or information about the

[3] This is counting only NUSAF2 beneficiaries in the communities included in the evaluation sample (see Section 3). The overall number of beneficiaries as part of NUSAF2 is much larger.

quality of a project alone is not sufficient. Rather, the combination of training on how to identify issues and report problems with information on the performance of projects leads to large improvements in outcomes from local development projects, in particular household assets.

We explore mechanisms for the observed impacts and find that the training and information increased community monitoring of the projects and cooperation among community members. The results are consistent with differential decapitalization between the communities offered the training and scorecard interventions and other groups. People report spending more time visiting and monitoring projects and making complaints to various levels of government. Individuals also report modest increases in the ability of communities to solve collective action problems and in trust in the central government. Qualitative work conducted ex post also suggests that the training and scorecard intervention induced some communities to take better care of the animals they had received, in addition to making complaints to local leaders. However, we do not have direct evidence that public officials delivered additional outputs based on citizen's complaints.

During a survey conducted before the experiment, we asked local leaders to identify areas near them that they thought had more corruption or mismanagement issues. We conduct heterogeneity analysis using these responses. We find that program impacts are concentrated in local areas that officials report as being more likely to be corrupt or mismanaged. We do not find spillovers across communities on our outcomes of interest, but we do find increased rates of monitoring of other projects or government services within treatment communities, suggesting the impacts observed here could expand to other public investments in treated communities.

An active body of research seeks to identify the most cost-effective approaches to reduce corruption and improve management of development projects. A recent systematic review by

Molina et al. (2017) finds that monitoring by communities can improve health services, though the evidence is limited. Research on the impact of community-based monitoring can be broadly divided into two types of interventions. The first involves providing trainings for communities to learn to identify issues on local development projects and how to act on them. The second involves providing information to communities on the quality or process of local development projects.

The evidence for the first type of intervention is limited. In one of the few studies on the provision of social accountability training to communities, Björkman and Svensson (2009, 2010) experimentally tested a program that combines information on the quality of providers and two half-day trainings to communities to improve provision of health care in Uganda. They find communities receiving this combined intervention monitored providers more, and these providers increased their effort levels. This led to reductions in child mortality and increased child weight. Björkman Nyquist, De Walque, and Svensson (2017) find that these results were sustained four years after the program. They also introduced another treatment arm with training only (on community monitoring), but their findings suggest that this was not enough to lead to sustained changes in the communities. Our results are consistent with those in Björkman Nyquist, De Walque, and Svensson (2017), in the sense that we show that it is the combination of training and information that leads to improvements in household assets. They did not, however, have an information-only treatment.

Evidence for providing information to communities is a bit more developed, though the results obtained thus far are mixed. In a well-known experiment, Olken (2007) tested the effect of dramatically increasing top-down audit rates and encouraging citizen monitoring of road projects in Indonesia. The community monitoring was done through accountability meetings,

where local leaders explained how funds were used. Communities received no other trainings or support to monitor that spending. Olken found significant decreases in leakages from the audits, but no effects from the community monitoring. Andrabi, Das, and Khwaja (2017) randomly provided report cards on school performance to communities in Pakistan. They found the report cards led to increases in test scores and enrollment and decreases in school fees. Banerjee et al. (2010) conducted a randomized evaluation of a program that tested whether community-created scorecards could lead to increased community participation in child education in India. They found this program had little impact. In another study, however, Banerjee et al. (2018) mailed information on a rice distribution program in Indonesia to inform households about the program, and find beneficiaries received significantly more rice. Finally, Barr et al. (2012) tested community-created scorecards on school performance in Uganda. Their findings indicate that the use of the scorecards increased student test scores and decreased teacher absenteeism. These varied results suggest that providing information can lead to improved service delivery, but information alone may not be enough, and the mechanisms are not yet well understood.

Our contributions to this literature are as follows. First, we provide evidence that social accountability training and information on project performance can empower communities to improve the public investments they receive. Our results suggest that project quality information or accountability training alone is not sufficient to improve services in a low-capacity environment; instead, both interventions need to be used together.

Second, these interventions were part of a large-scale, government-run program managed by the Inspectorate of Government and implemented in cooperation with local civil society organizations. As such, the scope, delivery mechanism, and scale of the program make it particularly relevant for learning about policy effectiveness (Muralidharan and Niehaus, 2017).

There is particularly little empirical evidence on the effectiveness of promoting social accountability in the context of large-scale national programs (Devarajan et al., 2011). Recent evidence on the differences in approach and impact of interventions by governments, NGOs, and small tightly designed experiments has led to concerns about external validity. Peters, Langbein and Roberts (2018) review 54 RCTs and find that almost two thirds are run with NGOs or researcher-managed interventions. Bold et al. (2018) find large impacts from an NGO-run intervention in schools in Kenya, while the same intervention run by the government has no impacts. Our results show that large-scale, government-led versions of social accountability programs can be effective.

The paper also illustrates how social accountability training and information interventions can be adapted and analyzed in the context of community-driven development or asset transfer programs that are delivering new services to communities. This is relevant given the large amount of resources committed to this type of interventions and their weak of effects on governance (for reviews, see White et al., 2018, Wong and Guggenheim, 2018). Social accountability interventions have traditionally focused on existing health and education services. As such, audits can be performed on these pre-existing services, and information interventions can then be based on these audits. This is not possible in our context since we analyze a program that had not previously delivered outputs to study communities. We first need these new services to be delivered before any audits can be conducted. In addition, the social accountability training was designed to potentially improve both the quality of the outputs delivered, and post-delivery monitoring. Therefore, we conduct physical project assessments after the social accountability training and the delivery of project output. The measure of the quality and quantity of project

output held by communities provides us with a first measure of training effectiveness. It is also the basis on which the information intervention is later implemented.

The results of this experiment suggest that low-income citizens can successfully obtain better outcomes from local development projects, when empowered with both proper skills and information. Large-scale, government-led versions of social accountability programs can increase the returns on investments in local development projects and improve citizen engagement. This can happen when social accountability training is combined with information about performance of local development projects. The effects can be especially strong in areas where local service delivery is particularly poor. Recent calls by international organizations for greater accountability is leading some to argue for reducing investments in areas where corruption and mismanagement can be high. Our results suggest that programs can instead implement a community-based monitoring approach to decrease the scope for corruption.

The remainder of this paper proceeds as follows. In the next section, we describe the NUSAF2 program, training and scorecard interventions. In section 3 we present the experimental design. In section 4 we present the data. We examine the results in section 5. Section 6 then concludes with a discussion of the implications of this work and a cost-benefit analysis.

2 The NUSAF2 program and interventions

NUSAF2 was a large-scale, community-driven development program implemented by the Office of the Prime Minister (OPM) in coordination with local, sub-county, and district authorities, with $135 million funding from the World Bank and the UK's Department for International Development (DFID) to the government of Uganda. We present a simple representation of the various levels of government in Uganda in the context of NUSAF2 in Figure 1.

Uganda is a small, landlocked country in east Africa. It is poor but has a stable and growing economy. Uganda, like many developing countries, faces significant challenges with service delivery. For example, though lowering child mortality and increasing rates of primary school enrollment are both major goals of the government, both of these measures of service delivery are poor (Bold and Svensson, 2013). Low-quality services can obviously be related to a lack of funding for programs, but even when money is available, service provision can also be a problem. Hard data on the sources of these issues are rare, though corruption and mismanagement by officials or service providers, as well as citizens' behaviors are often blamed.

NUSAF2 targeted villages in the poorest and least developed northern half of the country. As part of the program, communities were invited to formulate projects and submit proposals to project offices based in the sub-counties.[4] This process was done through the community driven development (CDD) model to increase local buy-in of development projects. Members of a community would gather, generally with support from a facilitator hired by the government and decide jointly on what type of project to apply for. The communities were responsible for developing the proposals and budgets, though local leaders would sometimes be involved.[5] Once approved by the sub-county, the proposals were then passed to the district, which assessed the feasibility of the projects before passing them on to OPM for final approval and funding. The submitted projects fell under three categories: (i) public works, (ii) livelihood investment, and (iii) infrastructure rehabilitation.

[4] "Community" refers to either a village or a collection of villages that come together to propose a NUSAF2 project. They are thus not legal designations but are official designations under NUSAF2. A village generally cannot receive more than one project.
[5] While very common in development programming, there is little evidence on how well CDD programs work relative to other policy instruments. The process of project and group formation is relatively opaque. Who is selected to be a beneficiery is left to the communities, and the process can vary across communities. It is possible that corruption may occur before the program has even been implemented if local elites or government officials hand-pick certain people through their social networks. The risks of elite capture have been analyzed in the literature on the targeting of social programs. As we cannot observe this well in the context of NUSAF2, we focus our analysis on potential issues that can arise during project implementation, and on how training and information treatments can improve project quality and development outcomes.

Once projects were approved by OPM, funds were transferred to communities, which managed the projects themselves through a variety of committees. Community Project Management and Community Procurement Committees were responsible for the delivery of the selected projects. Community Social Accountability Committees were created to oversee and monitor project progress and provide oversight within the community. Sub-county and district authorities were then expected to undertake monitoring and supervision in coordination with NUSAF2 project staff.

A highly decentralized project like NUSAF2 can create a range of transparency and accountability challenges.[6] Some concerns include that community and government officials may potentially misuse or divert funds from community projects. Anecdotal evidence from a previous phase of the program suggests some cases of misappropriation of funds by officials. If transparency is limited, communities may lose control over how money is spent. Officials may insist on low-quality suppliers for community projects, potentially expecting kickbacks. Community elites may try to engage in similar behavior to attempt to manage funds with little oversight or to induce fellow community members to hire low-quality suppliers.

At the same time, it is often impossible to separate corruption from general mismanagement of resources. Communities and local governments may simply not have the capacity to make optimal decisions, and so funds may be used inefficiently or ineffectively. It is also possible that there may be issues with collective action, where communities may fail to implement a project well because it is too difficult to organize community members to complete

[6] Evidence from Fisman and Gatti (2002) suggest that decentralization can actually *reduce* corruption. We do not take a position on whether decentralization in Uganda has increased or decreased corruption, only that a highly decentralized program can create a range of potential challenges.

the activities. Finally, beneficiaries themselves may simply fail to take sufficient care of public investments they receive.

To address these potential concerns, a Transparency, Accountability, and Anti-Corruption (TAAC) component was included in the design of the NUSAF2 project. We worked with the Inspectorate of Government to embed a randomized control trial as part of the component. In the seventh and eighth rounds of NUSAF2 funding (out of a total of 12 rounds), and after having been awarded funding for a specific project type, communities were trained on the details of project implementation and how to identify and prevent cases of corruption and mismanagement. The training was implemented by seven different CSOs across the broad north of Uganda,[7] which sent representatives to communities to implement detailed training on social accountability and community monitoring of NUSAF2 projects. The program also organized follow-up visits by CSO representatives to provide ongoing training and advise the communities on how to monitor implementation of NUSAF2 projects.

When the CSO trainers first entered a community, they organized community assemblies. In the first assembly, the trainers discussed the principles of social accountability and community monitoring and asked the community to elect representatives to add to an existing social accountability committee. The existing committees were generally considered to be untrained and poorly prepared to monitor issues in the project. The social accountability training was thus designed to give them new capacity. Members of the new committees made a public pledge to participate in the training program, undertake monitoring of the project on behalf of the

[7] Due to the size of the program, one civil society organization managed the implementation of the program but sub-contracted to seven individual CSOs that were present in the districts where the training was implemented.

community, and report back to the community. Approximately 5 people were selected to serve in social accountability committees in each community.[8]

The training provided background on social accountability and the NUSAF2 program, taught participants community-monitoring skills, and provided tools to monitor NUSAF2 projects. The training also provided hands-on skills in writing reports, giving feedback to the community, generating a community action plan, and applying monitoring skills to projects other than NUSAF2. The training gave special focus to encouraging communities to reach out and make complaints to the local and central governments, including the IG if necessary. People could contact the IG either by approaching a local office in their district or by texting a new national corruption hotline. A detailed description of the program components is presented in the appendix, including some of the visual training materials used for illiterate populations (Figures A1 and A2).[9] The training curriculum aimed to strengthen community monitoring, which was expected to lead to more complaints to public officials or improved cooperation to address issues at the local level. The training also included modules seeking to improve the procurement of project outputs through better selection of providers or improve interactions with local officials and service providers when project outputs are acquired.

Approximately six months after the mean completion date of these projects, from December 2015 through January 2016, we conducted an assessment of the quality and quantity

[8] While it is possible that local elites could have affected who was added to this committee, we did not observe this and our data suggests the selected people are not generally different than most members of the community. It is thus unlikely that local elites and local government officials participated in the trainings and felt scrutinized by the implementers.

[9] In addition to the main training treatment, an additional treatment was also attempted in a random sub-sample of communities. This additional treatment was supposed to increase incentives for individuals to monitor projects through non-monetary rewards. These took the form of pins provided to participants showing they served as community monitors. These individual incentives were low value. In addition, group rewards were considered for communities who completed the entire training, conducted the community monitoring and produced timely monthly reports. However, the implementing agency was not able to implement these group rewards. We compare the treatment effects between the different treatments and do not observe a meaningful difference in coefficients and significance. For the analysis presented here, we thus do not differentiate between the training treatments and instead present results of the pooled training treatment.

of the community projects. This was done through physical observations of project outputs. We then used this information to construct a score for the projects in each community. Details on the construction of these scores are presented in the appendix. In February 2016, individual community facilitators, trained by the research team but identifying themselves as representing the IG, went to communities to present these scores. The facilitators also provided communities a ranking of their performance, relative to other NUSAF2 communities in their district. The scorecard stated that their project was ranked X out of Y projects in the district based on their performance in the assessment. An example of a scorecard is presented in Figure 2.

To ensure comparability of scores, the scorecard was done only for livestock projects. (Due to operational issues, we also had to exclude the Karamoja region).[10] Treatment communities were presented summary information on the health of animals, animal productivity, assistance from the district veterinary officer (who was supposed to assist communities with their animals but was not always present), and a constructed "value for money" score that was calculated by multiplying the number of animals received by the productivity score of all the animals, divided by the total money received for the project.

During the dissemination of the scorecards, the communities were invited to discuss the results. This discussion was supported by the community facilitator and included opening remarks from community leaders and a speech introducing the goals of the meeting. The scorecard results were then announced, with each component of the score fully explained. The meeting ended with a discussion about how communities could use the results of the score to improve service delivery and accountability in the community. Some of suggested community

[10] The focus on livestock means that the information treatment was conducted only in projects that were a private good, as opposed to infrastructure projects that were a public good. We provide evidence below that the training treatment had similar impacts in livestock and other project types, but we do not have direct evidence on whether project quality information could have led to subsequent improvements in public good projects.

actions that were discussed during these meetings included: (1) voicing concerns to the sub-county and district leadership; (2) participating actively in the community projects; (3) voting for local politicians whom they believe can best help the community develop; (4) selecting the best possible project leaders and monitoring them closely; and (5) working together as a community to resolve issues whenever they can. These potential actions point to some of the mechanisms through which the scorecard intervention could subsequently impact project outcomes, namely (i) stronger community monitoring, (ii) more complaints to local officials and (iii) improved cooperation among local communities to resolve issues.

The facilitator brought to each community five copies of the scorecard in English and five copies in the local language, a number line to graphically show the ranking of the community project, and sodas and soap as gifts to participants. Once the facilitators left, they did not return to the community.

The training intervention we study here was based on a well-defined curriculum that was directly relevant for projects being implemented in communities. The training intensity was relatively long compared to other studies cited above. The scorecard information was also tailored to the projects and meant to encourage specific action by communities and presented direct comparisons to other communities in their area.

3 Experimental design

Due to the large size of the NUSAF2 program, it was implemented in twelve rounds over five years. Working with the IG, we were given a list of all NUSAF2 projects to be funded in the seventh and eighth rounds and randomized which communities would be given the social accountability and community monitoring training. The randomization of the social

accountability training and scorecard treatment was done in Stata. Due to the limited amount of administrative data from the government that had been digitized, we were only able to observe the location, budget, and classification of projects (whether public works, livelihood investments, or infrastructure rehabilitation projects). The communities' choice of project type was based on an endogenous process that we were not able to observe. Note that every community in our sample received a NUSAF2 project, and that the choice of project took place before randomization of the social accountability training and scorecard treatment. As the interventions we study here were randomized across projects, and project types are well balanced for each treatment and control condition, the type of project chosen by a community does not bias our inference of impacts from the training and information treatments.

The timeline for this study is presented in Figure 3. An initial survey of local officials, discussed below, was conducted in early 2013. In November 2013, we received the list of NUSAF2 projects chosen by communities and selected for funding as part of the seventh and eighth funding rounds. We randomly assigned communities into social accountability training treatment or control in January 2014, with the NUSAF2 program and social accountability trainings beginning in June 2014. In December 2014, 80% of the funds were distributed, with the other 20% funded in the preceding six months.[11] We conducted the project quality assessment from December 2015 to early February 2016. From this assessment we constructed the project quality information scorecard and randomized communities to receive the scorecard intervention in February. We then distributed the scorecards from February to March 2016. Six months after the assessment, in June to July 2016, we completed the final household survey. The final

[11] Most initial outputs were delivered to communities by October 2015, though some complementary outputs were delivered later, and communities continued to receive government services that may affect the quality of their project after that.

household data collection was done on a rolling basis to coincide with the timing of the project assessment and ensure that communities were visited on a consistent timeline.

The design and number of projects by type and treatment status for the social accountability training intervention are presented in Table 1. A total of 940 projects were included in the sample. However, our main outcomes are not easily comparable across each of the project types. In the project types with the smallest number of communities, we were unable to create a reliable index of outcomes from the first project assessment, and so we focus the analysis on the most common project types: enterprise development, fencing, livestock[12], road and housing construction, and tree planting. This reduces our sample to 895 projects.[13]

In Table 2 we present the information scorecard design. As described previously, we developed and delivered the project quality information scorecard only to communities with livestock projects to improve comparability. Due to operational difficulties, we did not include the northeast part of the country, the Karamoja region, with 61 communities. A total of 574 communities are thus included in the sample. The end design is a 2x2 that includes both social accountability training treatment and control communities.

The NUSAF2 program and the social accountability training were implemented across the broad north of Uganda. We present a map of training treatment intensity in Figure A3. The figure shows the number of NUSAF2 communities that received training by parish across the

[12] Livestock projects included cattle, sheep or goats. The livestock project sub-types are also balanced across treatment and control groups.

[13] Because we had information limited to the broad project type (whether (i) public works, (ii) livelihood investment, and (iii) infrastructure rehabilitation) before selecting communities for the social accountability training treatment, we were not able to pre-drop specific project types that were implemented in numbers too small to allow for reasonable comparison of project quality, as is commonly done in similar experiments. We instead drop them in our analysis here. As the number of such projects is small (less than 5% of the sample in total), and given that we target all the projects delivered by NUSAF2 in two funding tranches, this post-dropping does not affect internal validity.

entire sample.[14] In some areas, there is a high concentration of projects, but for the most part they are distributed across the broad region. We also look at spillovers at the sub-county level to test if the number of treated projects within a local area affects outcomes for the control group.

Before data were analyzed, all the outcomes were pre-registered with the American Economic Association registration system, number AEARCTR-0001115. The main outcomes of interest are the quality of the NUSAF2 project[15] and household assets. We analyze potential mechanisms, such as whether accountability training and project quality information scorecards affected community monitoring, complaints to public officials and cooperation to address issues at the local level. Our secondary effects of interest are whether the program changed individuals' perceptions of the legitimacy of local and central government, and whether there were spillovers to other programs in communities. The asset indicators include the number of cattle owned by the household as well as an index of total household assets. Cattle is highly relevant as it is a direct outcome of the most prevalent livestock projects and one of the most common way households store wealth in the area studied. The index of total household assets includes cattle and other livestock (such as goats and sheep), as well as household durables. We explore these effects for all projects but do not expect animal ownership to change in the non-livestock projects. We therefore constrain some analysis to livestock projects only.[16]

While we were able to confirm that all of the selected communities received training, and that training was of satisfactory quality overall, there were delays in training some communities.

[14] Administrative units in Uganda, from largest to smallest, go from the central government to the district, then sub-county, parish, and village. We present the intensity of projects by parish as it is a medium level of administration and best displays the intensities across the area.

[15] We describe in the next section and in appendix in Tables A1 and A2 the construction of this indicator.

[16] AEARCTR-0001115 describes the training experiment and contains pre-analysis plans that list the main outcomes and intermediary outcomes of interest covered in this paper. As the paper makes clear, the training and scorecard interventions cannot be analyzed independently, hence we report the results of the 2x2 experiment based on the original set of pre-specified outcomes and analysis in AEARCTR-0001115. The scorecard experiment is further detailed in AEARCTR-0003674, which outlines additional analysis beyond the scope of the original design.

The expectation was that communities would receive training either before or within a few months of receiving the NUSAF2 project funds. However, there are three reasons why this did not always happen. First, training implementers had limited information from the NUSAF2 program office about the timing of fund disbursements. Second, funds went from the central government to the districts before going to communities, and there was little information from the districts about their fund disbursement schedule. These two issues meant that precisely timing the training was very difficult in practice. Finally, the local CSOs often had difficulties organizing their activities to implement the training on time, and so delivered training later than originally planned in some cases.

Soon after the trainings were completed, we conducted a short process evaluation in a randomly selected 96 projects to determine when funds were received relative to when the trainings were conducted. We found that 17 projects received their training after they started using their funds, with 11 receiving training within two weeks of using their funds. Four projects (4.2% of the randomly selected sample) began using their funds at least a month before they received training. We consider this late treatment to be non-compliance. Given the low rate of late trainings, we do not make corrections for non-compliance and so focus on the intention-to-treat (ITT) estimates.

4 Data, empirical specifications, and balance

4.1 Data

The data for the analysis presented here come from several sources. Before the program began, we were given limited administrative data on which projects were to be funded by NUSAF2 in

each community. From this list, we obtained information on the location, budget, and project types.

We conducted a survey of local officials between January and March 2013 in which we included all 45 districts and 485 sub-counties in areas where NUSAF2 operated. Sub-county officials interviewed in the survey include elected and appointed officials, as well as local NUSAF2 officers. We were interested in obtaining information on levels of corruption or mismanagement at the local level. To measure this, we asked each respondent the following question: "In your personal opinion, within your district, which sub-county has the biggest problem with corruption?" We then counted the number of times a sub-county was mentioned. Of the sub-counties in the sample, 47% were never mentioned by an official and 20% were mentioned only once. We created an indicator if the sub-county was mentioned more than once.[17]

As mentioned above, outcome data were obtained from two separate surveys: first, a project quality assessment captures effects on community project outputs, and second, a survey of individual households conducted six months later that captures effects on the households. In both survey rounds, enumerators were blind as to the treatment status of the communities.

The first source of follow-up data collected is a project assessment conducted between December 2015 and February 2016. The project quality assessment includes observations of community projects by a team of enumerators. For projects with a single output (e.g., a staff house or a borehole), enumerators directly observed characteristics of the output. For livelihood support projects where outputs were distributed to beneficiaries, a sample of individuals was

[17] It is possible that communities select project types based on local prevalence of corruption. We in fact observe this. Communities that are in areas cited as corrupt choose livestock projects 58% of the time, while those in areas not cited as corrupt choose livestock 70% of the time. Note that the randomization of treatments occurred after project choice, so that these descriptive patterns do not affect the internal validity of the estimates.

randomly drawn from the list of beneficiaries and beneficiary-level outputs were observed. For example, for livestock projects, a sample of individuals was selected, and enumerators visited the sampled beneficiaries to observe the animals provided by the project. The project assessment data allow for the measurement of the quality and quantity of project outputs, as well as intermediary outcomes capturing underlying mechanisms through which the training could affect project outputs. For each domain, the project assessment allows us to capture a range of variables, which can later be aggregated into indices. The next sub-sections provide additional information on the main outcomes and intermediary outcomes to be tested and the indicators that were collected to measure them. The appendix provides tables with the full list of variables composing the indices (Tables A1 and A2).

The primary project-level outcome is a measure of a project overall score, which is composed of indices that measure the quality of the project and the quantity of outputs delivered. The project overall score is the main project-level outcome for the analysis. It is built as an interaction of a quality measure and a quantity measure. This allows us to account for situations in which a community received more output from a project but at lesser quality, and vice versa. The quality and quantity indices are also analyzed separately. As the quality and quantity indicators are created across different project types, the indices constructed are normalized within each project type to ensure comparability.[18]

Project quality is measured within each project type through direct observation of a range of attributes of the project output. For livestock, the project quality score is an additive index of whether the animal received was of the appropriate age, whether it was a local or improved breed of animal, whether the animal was productive when visited by the survey team, and whether the

[18] The indicators were normalized within each project type in the whole sample, by subtracting the mean and dividing by the standard deviation. See appendix and Table A.1.

animal displayed any signs of illness. For staff houses, the quality is measured in terms of how well the walls, roof, windows, doors, ceilings, and floors meet quality standards. For enterprise projects, quality is determined by whether individuals have access to materials, transportation, credit, labor, and markets. Road quality is measured by the material used in the construction. The quality of tree planting projects is determined by whether the seeds or seedlings are certified by the government or other NGOs.

The quantity measure captures the outputs of the community project. It is determined by the number of animals received, length and height of the building constructed, number of people engaged in the enterprise, length of the road constructed, and the number of trees planted. These measures are obtained from direct observations of the outputs by enumerators at the time of the project assessment. In cases where the output could not be observed, the quantity measure takes a value of zero. This happens for livestock projects, for example, when the livestock have died or are otherwise missing at the time of the follow-up project assessment. We provide the full list of quality and quantity indicators in the appendix. In addition, Figures A4 and A5 in the appendix illustrate how some project assessments were conducted in practice.

The final indicator considered is whether the project could be located for the project assessment. When the survey team was unable to find a project during data collection, a research assistant was sent to confirm whether the project existed. In total, 23 of the projects, or 2.6% of the sample, could not be found by the survey team during any of the attempts at data collection and so were considered missing projects. At the end of the data collection, the IG was notified of these missing projects. The IG office sent a team to verify their existence, which reported that they had identified each of the missing projects and confirmed they had been operating. It is unclear how these projects should be considered in our analysis. Significant efforts were made

by the survey team to locate the projects and confirm their existence. In addition, the missing projects were livestock and enterprise projects, which can be hard to identify because most households had multiple animals and income-generating activities prior to the project. It is possible that communities did not declare these projects to the survey team. It is also possible that, when the IG team arrived to confirm the existence of the projects, some communities presented similar types of output as coming from NUSAF2, even though these outputs may have previously existed. For our analysis, we test whether the share of these missing projects varies between treatment and control. For our measures of quality, we code these projects as zeros. Most importantly, the results are also robust when treating these projects as survey attrition and dropping them from the analysis entirely.

In addition to the primary project-level outcomes, the project assessment also measures three sets of intermediary outcomes that capture the main underlying mechanisms through which the training can explain changes in final outcomes. These include (i) community monitoring, (ii) the procurement and contracting process, and (iii) community interactions with local leaders. These three domains relate to some of the key areas covered by the social accountability training curriculum. Intermediary outcomes include indicators of community monitoring, such as an index of the intensity of project community monitoring, and an index of the intensity of social accountability committee (SAC) project monitoring. Indicators on the procurement and contracting process include an index of challenges faced by communities in the procurement process, an index of satisfaction with suppliers of goods and materials, and whether the community hired a contractor. For communities that did hire a contractor, indicators also include an index of challenges faced by communities in the contracting process and an index of satisfaction with the contractor. Finally, the third main domain for intermediary outcomes

captures interactions between communities and local officials. This domain includes indicators of whether a payment was made to a district official or staff, and an index of satisfaction with the sub-county NUSAF2 official and district veterinarian officer.

The second source of follow-up data is an endline survey conducted with households in the sample communities in June and July 2016. The sample surveyed was a random selection of NUSAF2 communities, which are made of 10-15 individuals that come together to form a project.[19] Eight people per community were surveyed. These include the two chairpersons of the executive committees in the project, two members of the original community social accountability committee, and two regular members. In the social accountability training treatment group, two members from the expanded community accountability committee (called the CMG) were also surveyed to assess how their profile differed from other members, but they are not included in the sample used for estimation of treatment effects as these individuals are not surveyed in control communities. In the social accountability control group, the CMG does not exist, and so two additional regular members were surveyed instead. The sample used in the analysis is thus a stratified sample composed of eight individuals in social accountability training control communities and six in social accountability training treatment communities.[20]

The data from the household survey contains assets, including animals and household durables; whether the individuals had made complaints to local leaders about their NUSAF2

[19] See footnote 3. Also note that the definition of communities is the same as the one used by Blattman et al. (2014) when analyzing another type of intervention delivered as part of NUSAF.

[20] Note that, as the individuals surveyed were selected randomly and were stratified by type of NUSAF2 project members (depending on whether they were member of committees or not), they are representative of the communities. We include project leaders as they are an important sub-set of beneficiaries, since most beneficiaries were invited to participate in one of the project committees. As a robustness check, we include controls for respondent role and do not find any differences in the results. We also check for heterogeneity in outcomes between leaders and general members and do not find differences either. Finally, we weight people based on the inverse probability of them being selected, as well as randomly drop two general members from the control communities, and find the same results. The inclusion of two additional community members in the control group does not bias our estimates.

project or other projects in the community; and the individuals' level of trust in local leaders. The descriptive statistics for the project assessment and household data collections are presented in Table A3. The description is separated by whether data were collected at the project or household level. While NUSAF2 targeted very low-income households, most had livestock in their home, with the mean household having 2.45 cattle at the endline.[21]

Our main outcome focuses on assets, specifically cattle, for several reasons. First, it is the most objective measure that we could identify. Second, it is generally used by researchers as a proxy for wealth in low-income settings. Finally, increasing the number of animals in communities was the expressed purpose of NUSAF2 livestock projects.[22] As mentioned above, we also present results for a livestock index and household asset index capturing a broader range of assets, including other livestock such as goats and sheep.

The sample size for the household survey was determined to provide the highest statistical power given a fixed budget. The intra-cluster correlation (ICC) for the main outcome of interest, number of cattle, is 0.045. For the scorecard sample, which includes 574 clusters, the minimum detectible effect (MDE) size is below 0.10 standard deviations. For total assets, the ICC is 0.35 and so the MDE is approximately 0.15 standard deviations.

[21] As part of a separate experiment, the enumeration teams were randomly assigned to villages during the endline data collection. This was done to test for enumerator effects on reported household characteristics and outcomes. There is no or very little enumerator bias introduced on the main outcomes of interest, especially number of animals. While the experiment is not able to directly test for Hawthorne effects, the lack of enumerator bias and the fact that the enumeration team was separate from the implementation team reduces the likelihood of such issues impacting the main results.

[22] While another good indicator would have been the quality of livestock present, as we use in the project assessment survey, quantity captures quality in one critical way: fewer animals have died, which is one of the biggest issues with animal quality in Uganda.

4.2 Empirical specifications

We start by estimating the impact of the training intervention on project-level outcomes measured through the project assessment. This analysis is done on data collected before the scorecard intervention was implemented. In this case, we estimate the intention-to-treat (ITT) OLS regression model:

$$Y_i = \gamma_0 + \gamma_1 Training + \omega R + v_i \qquad (1)$$

where i refers to a project and Y_i is the outcome of interest. *Training* is whether a community was randomly selected to receive the social accountability training. R is a matrix of sub-county dummies and v_i is the error term. This specification provides an estimate of the overall effect of training at the project level. We do not consider the effect of the scorecard in this specification, since the scorecard intervention occurred after the project assessment.

We then present estimates of the impact of the training and scorecard interventions on household-level outcomes measured through the follow-up household survey. We run the following intention to treat (ITT) OLS regression model:

$$Y_i = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_3 T_3 + \varphi R + \varepsilon_i \qquad (2)$$

where i refers to a household and Y_i is the outcome of interest. T_1 is whether a community was randomly selected to receive the social accountability training treatment only, T_2 is the scorecard treatment only, and T_3 refers to communities assigned to both social accountability trainings and

scorecard distribution. R is a matrix of sub-county dummies and ε_i is the error term.[23] The coefficient β_1 thus presents the impact of the social accountability training treatment only, β_2 the impact of the scorecard treatment only, and β_3 the impact of the combined social accountability training and scorecard treatment. For household-level outcomes, we cluster the standard errors at the project level, which is the level of randomization. Note that this specification provides an estimate of the effect of training only (β_1), which is different than the overall effect of training at the project level (γ_1 in equation (1)). Besides estimating β_1, β_2 and β_3, we also present results for tests of $\beta_1=\beta_2$, $\beta_1=\beta_3$ and $\beta_2=\beta_3$.

Note that we have two main final outcomes: project score (measured during the assessment) and household assets (measured at the household endline). Both of these outcomes represent indices of family of outcomes. To further explore the potential mechanisms, we discuss impacts on indices of community monitoring and reporting to government officials as reported by respondents. We present some analysis on individual components to test for mechanisms, which we consider to be exploratory analyses. We end by looking at important heterogeneities in treatment and local spillovers.

4.3 Balance tests

Table 3 presents balance tests for the estimation of the impact of training on project outcomes in the full sample (panel A), and for the estimation of impacts of the training and scorecard treatments in the scorecard experiment sample (livestock projects, panel B). Due to the project

[23] In addition to this specification, we test for robustness by including additional controls. These include respondent role in the project (executive committee chairperson, member of original social accountability committee, or regular member), and demographics collected at endline. We do not find any difference in our outcomes when including these controls.

timeline and funding, a full baseline with communities was not feasible.[24] We have four indicators that were available before the beginning of the NUSAF2 projects in the sample: the type of project, the amount of money approved per community, when the program grants were received, and the level of corruption and mismanagement in the areas where the communities are located. We also present tests for whether the livestock project provided cattle (as opposed to goats or sheep), whether randomly drawn respondent in the household survey was a man, whether that person could write or read, and the distance from the respondent's household to the sub-county headquarters. We include these last four measures because we believe they are not likely to have changed due to the program and so reflect the characteristics of the communities before the social accountability training treatment.

For the training experiment sample (Table 3, Panel A), we do not find a statistically or economically significant differences between the social accountability training treatment and control groups for indicators at the project level. There is no difference in the likelihood of the project being livestock, for livestock projects to provide cattle, in project funding, or in the date when the funding was received in the communities. Turning to individual characteristics of community members, there is a difference in whether sampled participants were men, whether they could write and the distance to sub-county headquarters. While these differences are statistically significant, they are relatively small in magnitude. In addition to the specifications in the paper, we implement specifications controlling for these indicators and do not find any differences in the main results.

[24] For a discussion of when a baseline is not necessary, see Muralidharan and Niehaus (2017). They argue that baselines in large-scale experiments with governments can increase the risk of the research not being completed and, with large enough sample sizes, are not strictly necessary. We reached the same conclusion in this study during the design phase, and prioritized data collection investments in large-scale follow-up surveys.

The livestock projects in the scorecard sample are well-balanced (Table 3, Panel B). There is a small difference in the amount of project funding and in the share of respondents able to write in communities assigned to the combined treatment, but the differences are again small in magnitude (1% of the control mean for funding, 8% for literacy, and 8% for distance). Overall, we consider that the characteristics of the communities and the people within the communities are generally balanced due to randomization. Where imbalances are found, they are of small magnitude and we control for them in robustness checks that show they are unlikely to affect our main results.

5 Results

5.1 Social accountability training impact on project outputs

The impacts for the main project-level outcomes are presented in Table 4. These include the overall score for each of the NUSAF2 projects in the sample (columns 1 and 2), which is created by multiplying the project quality score (columns 3 and 4) and quantity score (columns 5 and 6). We also look at whether the project could not be located (columns 7 and 8). These indicators are from the project assessment survey and project-level estimates are obtained based on equation (1). The indicators are standardized, as discussed previously.

Odd-numbered columns contain results for all project types in the sample. We find a small positive impact of 0.119 standard deviations on the overall project score (significant at 10% level). This effect is mostly driven by the quantity indicator (column 5), and not by whether the project could not be located (column 7). The results suggest that the training led to an increase in the quantity of outputs delivered by projects by approximately 0.185 standard deviations (significant at 1% level). The point estimate of the project quality score is positive,

but not significant. In appendix Table A4, we show that results are consistent when top-coding observations above the 99th percentile.[25]

Even-numbered columns in Table 4 report outcomes for interacting treatment with whether the NUSAF2 project was non-livestock. We look at this difference specifically as we are interested in whether the results are being driven by a specific project type. As most project types are a small portion of our total sample, we are only able to look at livestock projects, which are about two-thirds of the total sample. Livestock projects are also the project type most likely to directly lead to welfare impacts at the household level and the focus of the scorecard experiment, as discussed further below. The coefficients for treatment effects remain about the same size. However, most likely due to decreases in power, the project overall score is not significant at the 10% level, though the quantity score is still significant at the 5% level and close in magnitude to the non-interacted results. For livestock projects, the impacts on the quantity scores of 0.167 standard deviations is approximately equivalent to an extra 0.9 heads of cattle per community.[26] None of the interaction terms in Table 4 are significant. We conclude there is likely no large difference between the impact of the program on livestock and other projects.

To further explore what is driving the impacts on the project-level indices, we present in Table A5 the components of livestock-only projects scores aggregated across animals at the beneficiary level. Consistent with the increase in the quantity scores in livestock projects, we find a decrease in whether the animal was not present during the assessment and reported to the team as dead, stolen, or sold by 4.6 percentage points. This is driven primarily by dead animals, and suggests slower decapitalization of project outputs in the training treatment group.

[25] A randomized inference test produces results similar to the OLS results, and so we only present the results of the OLS specification.
[26] On average, projects providing cattled delivered 13.6 heads of cattle per community.

Consistent with the lack of significant impacts on project quality, we find no statistically significant impacts on the age of the animal when it was purchased by the community, the breed of the animal, whether the animal was deemed productive. We only find a small improvement in the health of the animals.[27]

We conclude that the social accountability training led to modest improvements in the outputs delivered by local projects, driven by an increase in the quantity of outputs, with more limited effects on the quality of these outputs.

5.2 Impacts on household assets

Six months after the initial project assessment, we conducted an additional household survey in the livestock sample communities to measure household-level assets. The household survey allows us to go beyond the measures of project outputs obtained from the project assessment and provide finer estimates of the impacts of the social accountability training and scorecard information treatment at the household level.

In Table 5, we present results for the main outcomes of interest from the household survey: the number of cattle held by the household, a livestock index and an index of total household asset ownership. The livestock index aggregates different types of livestock using tropical livestock units.[28] Note that we do not expect impacts on the number of household animals for any but the livestock projects. We prespecified a focus on this outcome as livestock projects represent 68% of the sample, and we believe these are the projects that are most likely to

[27] Note that the illness index is reweighted as 1 minus the mean number of illnesses, so the positive coefficient means fewer observed illnesses.

[28] Cattle and household asset outcomes were pre-specified. Results are provided for the livestock index as an additional robustness check. The livestock index = 0.7 * number of cattle + 0.2 * number of pigs + 0.1 * number of goats and sheep + 0.01 * number of poultry.

lead to direct changes at the household level. They are also the focus on the scorecard intervention. For this analysis, we restrict the sample to beneficiaries of livestock projects, i.e., those who were selected to receive animals from NUSAF2.[29]

The results from the household survey show that the combination of the social accountability training and scorecard treatments leads to impacts on household assets. The number of heads of cattle held by households who received both the social accountability training and project quality information scorecard interventions increase by 0.421, or approximately 19 percent relative to the control group.[30] This is a highly statistically significant effect of substantial economic magnitude. Results also show a significant treatment effect of 0.3 tropical livestock units in the livestock index (a 16 percent increase relative to the control group), and of 0.237 standard deviations in the total asset index.

Importantly, the training treatment or the scorecard treatment by themselves do not have significant impacts on cattle, livestock or assets at the household level. The pairwise tests of equality between the treatments show that it is the combination of treatments that drive observed impacts. We can reject that the effect of the combined treatment is equal to the effect of the training only (p-value of 0.09 for cattle, 0.06 for the livestock index and 0.04 for the total asset index). We can also reject that the effect of the training and scorecard treatment is equal to the effect of the scorecard treatment only (p-value of 0.08 for cattle, 0.03 for the livestock index and 0.06 for the total asset index).

[29] The endline survey was conducted on eight beneficiaries per community in the control group. In the treatment group, we included six beneficiaries, as well as two non-beneficiaries who were selected to join the community managements committee as part of the training intervention. We do not include non-beneficiaries in this analysis as we do not expect impacts from the treatment on household welfare. Table A15 in the appendix provide robustness check for randomly dropping two general members from the control communities (as in footnote 19).

[30] We also test for whether impacts are concentrated in communities that had the lowest scores (not shown) and do not find a relationship between the absolute score and the number of animals in households. The impact of the training and scorecard information appear to exist across the distribution of scores.

We interpret these results as showing that the scorecard treatment, without social accountability training, is not sufficient to improve assets at the household level. At the same time, while the training treatment led to a small increase in the overall project outputs delivered by local projects, this effect alone is not sufficient to lead to significant increases in the number of assets held by households at endline. The combination of treatment, however, leads to significant and large improvements in household assets.

5.3 How do midline and endline results compare?

We next summarize how the magnitude of the midline and endline results are consistent with each other. Projects providing cattle delivered on average 13.6 heads of cattle per community (see footnote 25), or 1.13 cattle per household for a community of 12 households. The midline effects can be re-expressed in cattle per household, i.e. in a scale similar to the one used for endline outcomes. Doing so, the 0.167 standard deviation increase in the project quantity score is equivalent to a difference of approximately 0.92 heads of cattle per community between the training and control groups at midline (up from a mean output of 13.6). With approximately 12 beneficiaries per project, there are hence 0.08 additional head of livestock per household between the training and control groups at midline. This illustrates that the effect of the training only is limited. At endline, there are 2.18 cattle per households in control, 2.3 in the training group and 2.6 in training with scorecard group, hence an additional 0.42 cattle per household between the scorecard and training and the control groups. This is equivalent to 4 heads of cattle per project, valued at $230 each.

In the next section, we discuss potential mechanisms explaining the differences between the midline and endline results. While we are not able to formally disentangle mechanisms, we

can suggest plausible pathways explaining the results. One potential explanation would be an additional injection of assets into these communities. However, we do not have evidence for this channel. Rather, the results are consistent with differential decapitalization between the communities assigned to the training and scorecard interventions and other communities. This differential decapitalization can be driven by better quality animal, improved collective action and communities taking better care of their animals, as well as better service provision from local officials, such as veterinary services.

5.4 Potential mechanisms: impacts on community monitoring, reporting and action

We now explore potential mechanisms that can contribute to explain the observed impacts. To recall, based on the content of the training and scorecard interventions, it was hypothesized that they would lead to stronger community monitoring, more complaints to local officials and improved cooperation among local communities to resolve issues.

We start by analyzing the impacts of the training on actions taken by communities to monitor their projects. In the project assessment survey, we created indices on the intensity of monitoring activities by the local accountability group that is present in all communities, as well as by the broader community. Table A6 (columns 1 and 2) documents the impacts on the social accountability training on these indices, based on the specification in equation (1) and the full sample including all project types. We find a small increase in the intensity of monitoring by the broader community, significant at the 10 percent level. However, we find very large and significant increases in the intensity of monitoring activities conducted by the social accountability committee group. This is consistent with the focus of the social accountability

training, showing it effectively increased project monitoring by the social accountability committees, and, to a less extent, by the broader community.

In Table 6, we explore the impacts of the training and scorecard interventions on communities reporting issues to officials at different levels of government. We use measures from the household survey conducted six months after the project assessment. We present an index capturing the intensity of reporting of issues as part of NUSAF2 projects (column 1), and its individual components (columns 2-5). Estimates are based on the specification in equation (2) and the scorecard experiment sample. The results show a significant increase in the number of reports in the combined treatment. The increase in reports is observed at all levels of government.[31] Complaints to the lowest level of government, LC1 and sub-county officials, increase by approximately 20-25 percent. Complaints to officials at a slightly higher level of local government, the district, increase by 48 percent. Finally, complaints to the central government through the IG increase by 150 percent.

The results show some impacts of the social accountability training only, and the scorecard treatment only on the number of issues reported by communities. Importantly, however, Table 6 suggests that the intensity of community monitoring and reporting of issues was stronger in communities assigned to receive both the social accountability training and scorecard treatments. Using the aggregate index, we can reject the hypothesis of equal treatment effects between the combined treatments and the training only (p-value = 0.006), while the difference between the combined treatment and scorecard treatment is marginally insignificant (p-value = 0.113).

[31] Table A7 in appendix document similar increases in reporting of issues when looking at all project types and only the social accountability training treatment.

In Table 6 (in columns 6-8), we further test for potential impacts on collective action at the community level. Column 6 captures a proxy index for collective action, based on two measures of the ability of communities to come together to solve problems and deliver public goods (in column 7 and 8). Results suggest a small effect of the combined treatment on whether respondents believe members of the community can come together to solve issues by themselves. While the increase in collective action is larger in the combined treatment than in the scorecard treatment (p-value=0.046), the difference between the training only treatment and the combined treatment is not significant. These results suggest that the combination of the social accountability training and project information scorecard may have affected the ability of communities to cooperate and resolve issues at the local level, but the effects are relatively small and are not larger than the effects of the training only.

In Table 7, we test whether treatment affected individuals' perceptions of the performance of the project leaders. We ask about general satisfaction with the quality of the project and management committee, respectively. We do not find meaningful changes in any of these measures. Similar results are found in appendix Table A10 for the full sample and only social accountability training treatment status.

We present results for additional potential mechanisms in appendix. Aside from strengthening community monitoring, it was originally hypothesized that the social accountability training could improve project-level outcomes by affecting how communities procured outputs for their projects, or by changing interactions with public officials when communities were in the process of obtaining project outputs. We do not find supportive evidence for these alternative mechanisms. This is consistent with the effects coming mostly from the combination of training and scorecard information. Table A8 documents impacts of the

training on the procurement and contracting process at the project level. We do not find statistically significant impacts on whether communities had issues with procurement or contracting, their satisfaction with suppliers, whether they hired a contractor, or their satisfaction with the contractor. We interpret these results as suggesting that the social accountability training did not affect the ways communities procured the outputs for their project. Table A9 documents the impact of the training on communities' interactions with bureaucrats, including whether communities report making payments to officials and are satisfied with technical officers providing services to their projects. This is again analyzed at the project level. We do not find impacts on whether communities made a payment to a district representative, their satisfaction with the local NUSAF2 coordinator, or their satisfaction with the local veterinarian officer.

Overall, community members made significant complaints to local officials and organized themselves more. We observe large and significant effects on monitoring of projects and complaints to officials. We do not observe differences in the way project outputs were procured, or in interactions with local officials. We interpret these results as mostly pointing to changes within the communities. The combination of training and scorecard information dissemination led to stronger community monitoring, contributed to the identification of issues, led to an increase in complaints to local officials, and induced an increase in communities' ability to address these issues themselves. This likely contributed to a slower decapitalization of project output in communities receiving both training and scorecard information.

There are of course several other mechanisms that could have contributed to the observed increase in household assets. In particular, the interventions could have changed information asymmetries by supporting communities in understanding better what was to be delivered to their communities and how that was to be done. The trainings could have also led to changes in

bargaining power by communities. For instance, a local newspaper reported on the arrest of a NUSAF2 official by the IG, instigated at the request of a treatment community.[32] However, we do not have direct evidence that public officials delivered additional outputs based on citizen's complaints. Qualitative work conducted ex post also suggests that the training and scorecard intervention induced some people to take better care of the animals they had received, besides making complaints to local leaders. This suggests that the results could be driven by livestock dying or disappearing less fast in the combined treatment group.

5.5 Impacts on trust in community leaders and government

We also analyze whether the program changed the way people view local and government officials. In Table 8 we present the results from asking respondents whether they thought their leaders acted in the interests of local communities. In columns 1 to 6 we look at the community leaders for the NUSAF2 program, the elected sub-county official, sub-county bureaucrats, the elected district official, the district bureaucrats, and the central government, respectively.

We do not find significant overall changes in how people perceive project leaders, or the sub-county and district elected and appointed officials. We do find a statistically significant increase in trust in the central government. This effect is small in magnitude compared to the level of trust in the control group. We find a similar result in appendix Table A11 on the full sample, looking only at social accountability training treatment effects. We believe this effect could reflect the increased visibility of the IG, the agency from the central government that managed the interventions delivered to these communities. It could also be due to the fact that the training and scorecard treatment led to increased interactions between community members

[32] http://www.monitor.co.ug/News/National/Nusaf-officials-arrested-over-theft-of-more-than-Shs4O0m/688334-2704288-j4ahwnz/index.html

and officials at all levels of government, with particularly strong increases in complaints to the central government, as shown in Table 6.[33]

5.6 Heterogeneities by local levels of corruption

To determine which communities in our sample had the largest issues with corruption and mismanagement, we conducted a survey of all local officials in the areas that would be part of the study before the start of the experiment. As described above, we asked officials to name the most corrupt or mismanaged sub-county in their district.[34] We then count the number of times a sub-county is listed and create an indicator of whether a given sub-county is in the top most cited sub-counties. If a sub-country is mentioned more than once we consider it to have high reported corruption. This is the case in 33% of the sample sub-counties.

In Table 9 we present the results of dividing the sample by this indicator and testing for the impacts on the number of cattle per household. The impact of the treatments is concentrated entirely in communities in the sub-counties noted by local officials as most corrupt or mismanaged. The social accountability training treatment indicator is marginally not statistically significant, while the interaction between social accountability training and project information scorecard is significant at the 1% level and very large.[35] Households in areas that are reported more corrupt or mismanaged that received both treatments have, on average, an additional 1.41 animals. This is an increase of 58% over the control group. This result suggests there is substantial heterogeneity in the effects from the interventions, and so there could be large

[33] Overall, trust in leaders and officials tend to be marginally lower in the combined treatment than in the training only treatment. This is particularly the case for trust in local officials at the project, sub-country and district levels, for which we can reject equality in the level of trust between the training and combined treatments.
[34] A district is composed of approximately five sub-counties.
[35] Rather than splitting the sample, the same results can be obtained in an interacted specification. In this case, the coefficient of the interaction term between ”Training and Scorecard” and ”High reported corruption” is significant at the 5 percent level.

benefits from targeting such a program to areas that have the biggest issues with corruption or mismanagement. We also test for differential mechanism effects (not shown) and find no difference in intermediary outcomes between areas of high or low reported corruption. We conclude that people increased project monitoring and made similar levels of complaints in these areas, but follow-up actions were most effective in the areas reported as more corrupt at the local level.

While we do not observe a large difference in control means across the high and low reported corruption groups in Table 9, we cannot rule out that this measure could also be correlated with other community characteristics, including performance of local government and overall poverty levels. We compare the results of the reporting of officials and the scores in the scorecards and find a significant relationship between the two. Communities in sub-counties that are most more likely to be corrupt have a lower absolute score of 2.39 points, out of an average of 70.85, significant at the less than 1% level. The results suggest that officials likely have very useful information on the level of corruption and mismanagement at the local level.

5.7 Spillovers and other heterogeneities

We conclude by testing whether the interventions induced spillovers at the local level. We first test for spillovers on community monitoring of other government project and services within communities. We then test for spillover across communities, which could capture the effects of higher-level officials shifting corrupt practices from areas with higher-intensity treatments to areas with lower-intensity treatments.

First, within communities, there could be effects from the treatments on other community projects not related to the NUSAF2 projects. In Table 10 we recreate the analysis on reporting of

issues conducted in Table 6, but for other community projects that are not related to NUSAF2. Like the results for NUSAF2-related projects, we find statistically significant and large effects on whether community members report issues. Reports from individuals about making complaints to officials at all levels about non-NUSAF2 projects (column 1) increase by 25 percent in the combined treatment group compared to the control group.[36] This is mostly driven by additional complaints to the IG, district officials and to a less extent sub-county officials. We cannot reject equality in these spillover effects across treatments, possibly due to the relatively smaller magnitude of the observed effects.

Second, while the randomization process for the selection of treatment communities was not stratified, it led to natural variation in the number of treated communities within sub-counties. We utilize this variation to look at the spillovers of treatment to control communities. To do this, we focus on the treatment group that received both the training and information treatments, and then calculate the total number of treated communities by sub-county, divided by the total number of projects in our sample. This provides an indicator of the intensity of treatment. Spillover effects could be positive if local officials feel pressure from communities and so improve all of their operations. They could also be negative if officials shift corruption or mismanagement from treatment to control communities. We present the results in Table A13. The size and significance of the coefficient for treatment is identical to that found in Table 5, (column 1). The coefficient for the intensity of treatment in a sub-county is large, but not significant. We do not detect observable spillovers from the program on control communities.

[36] When looking at the full sample of communities in appendix Table A12, we also find significant effects from the unconditional social accountability training treatment on reporting, although the effect sizes relative to a control group are generally small.

Overall, the findings suggest spillover within communities, but not across communities, which is broadly consistent with the intervention mostly having impacts at the local level by inducing populations to address issues found within communities.

We also look at heterogeneities by distance from individual beneficiaries' homes to the main sub-county office, by beneficiary sex, by beneficiary ability to read, by whether the beneficiary is connected to the local leader (lc1), by regular beneficiary (as opposed to other committee members/leaders), and by whether the livestock sub-projects delivered cattle.[37] The results of these tests are presented in appendix Table A14. Overall, we do not find significant variation in outcomes on the individual characteristics such as sex, connection to local leaders, or whether individuals are regular beneficiaries or not. There does appear to be some variation in outcomes depending on individuals' ability to read. In particular, the scorecard intervention may have been more effective among low-skilled individuals, who may otherwise not have been able to gather information on their own. Finally, and consistently with results so far, impacts on cattle ownership are driven by livestock sub-projects providing livestock. (Although the interaction term is not statistically significant, the coefficient for the scorecard and training group is only significant in the sub-group of households in livestock projects providing cattle).

6 Discussion

The impacts of combining social accountability training and scorecard information on household assets in communities that received livestock projects are quite large. At the time of the follow-

[37] The list of heterogeneity dimensions slightly deviates from the list included in the pre-analysis plan. For example, since the scorecard intervention could not be implemented in Karamoja, we cannot perform heterogeneity for this region. Heterogeneity by asset was not performed since impacts on household assets are documented. Heterogeneity by perceived importance of livestock or outside interference was not performed as there is not enough variation in these variables (nearly 95% of respondent state that livestock is very important to them and few reported outside interference in the purchase of material). In addition, we also show heterogeneity by type of livestock project, as per the useful suggestion of a reviewer.

up survey, we estimate that there are over four additional cattle in communities that received both treatments. At the time of the program, cattle were valued at approximately 800,000 USH, or about $230 each. The social accountability and scorecard information thus led to approximately $97 worth of additional animals per household, or between $970 to $1,455 per community. However, the cost of the program was significant, given the geographic spread and relative intensity of the training. We estimate that the total cost of the program, measured by the amount paid to the CSOs that ran the trainings, was between $900 and $1,200 per community, depending on how costs are accounted.[38]

There are two points to keep in mind with this cost/benefit calculation. First, there is heterogeneity by local levels of corruption. The impacts we observe are concentrated in communities that were considered by local leaders as more corrupt or mismanaged: the effects are up to four times larger than the estimated average treatment effects. For communities that are particularly likely to be affected by corruption or mismanagement, the combination of social accountability training and information has especially large effects. Second, we assign zero value to livestock that households do not hold anymore. This may overestimate welfare impacts if households derived some values from livestock no longer held.

A final note should be made on the external validity of this study. The large-scale experiment we conduct has features that are often considered as increasing external validity, such as geographic scale and implementation through government agencies (Muralidharan and Niehaus, 2017). At the same time, like all research, context can matter for the results we obtain. Specifically, for the scorecard treatment we are studying a component of the NUSAF2 program which may be easier to monitor (did you get a livestock, is it dead or alive, is it productive, etc.),

[38] These costs reflect the time spent developing the material; training of the CSO representatives; transport, materials, and drinks for participants during the trainings; and scorecard dissemination.

while excluding other projects that may be more difficult to monitor, or require more coordination within the communities (infrastructure, roads, etc.). However, livestock was the largest component of the NUSAF2 program, and is a common intervention in asset-transfer and community driven development programs around the world.

Overall, we present evidence that increasing engagement in poor communities can produce higher returns from public investments. The social accountability training, combined with a project quality information scorecard intervention, resulted in individuals owning a significant number of additional animals. These effects appear to come from increased monitoring by communities, as well as an increase in the reporting of issues to officials from the local to the central government. We also find that the program led to some modest improvements in people's trust in the central government.

The results suggest a positive role and significant potential for programs that seek to promote citizen engagement and increase local populations' participation in the development process. This approach is becoming popular, with similar interventions being conducted in large-scale government programs in Liberia and Sierra Leone, as well as being expanded considerably in Uganda. We show that this approach is feasible, impactful, and, under some conditions, of good value. But it is clear that communities in this context need more than training on how to identify and report issues alone, or simple information about their project's performance. Rather, it is necessary to combine these interventions, especially in areas where citizen's interactions with government are difficult or not the norm.

References

Andrabi, T., J. Das, & A. Khwaja. 2017. "Report Cards: The Impact of Providing School and Child Test Scores on Educational Markets." *American Economic Review* 107 (6):1535-63.

Banerjee, A., Green, D., Green, J. & Pande, R. 2010. "Can voters be primed to choose better legislators? Experimental evidence from rural India." Mimeo.

Banerjee, A., R. Hanna, J. Kyle, B. Olken & S. Sumarto (2018). "Tangible information and citizen empowerment: identification cards and food subsidy programs in Indonesia." *Journal of Political Economy* 126(2).

Barr, A, F. Mungisha, P. Serneels and A. Zeitlin, 2012. "Information and collective action in the community monitoring of schools. Field and lab experimental evidence from Uganda". Mimeo.

Bertrand, M., Djankov, S., Hanna, R. & Mullainathan, S. 2007. "Obtaining a driver's license in India: An experimental approach to studying corruption." *Quarterly Journal of Economics*, 122, 1639-1676.

Björkman, M. & Svensson, J. 2009. "Power to the People: Evidence from a Randomized Field Experiment on Community-Based Monitoring in Uganda." *Quarterly Journal of Economics*, 124, 735-769.

Björkman, M. & Svensson, J., 2010. "When is community-based monitoring effective? Evidence from a randomized experiment in primary health in Uganda." *Journal of the European Economic Association* 8(2-3): 571-581.

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

Bold, Tessa, and Jakob Svensson. 2013. "Policies and Institutions for Effective Service Delivery: The Need of a Microeconomic and Micropolitical Approach." *Journal of African Economies* 22 (suppl_2):ii16-ii38.

Bold, Tessa, Deon Filmer, Gayle Martin, Ezequiel Molina, Brian Stacy, Christophe Rockmore, Jakob Svensson, and Waly Wane. 2017. "Enrollment without Learning: Teacher Effort, Knowledge, and Skill in Primary Schools in Africa." *Journal of Economic Perspectives* 31 (4):185-204. doi: doi: 10.1257/jep.31.4.185.

Bold, T., M. Kimenyi, G. Mwabu, A. Ng'ang'a, & J. Sandefur. 2018. "Experimental evidence on scaling up education reforms in Kenya." *Journal of Public Economics* 168:1-20.

Devarajan, S., S. Khemani, And M. Walton. 2011. *Civil society, public action and accountability in Africa*, The World Bank.

Fisman, R & Gatti, R. 2002. “Decentralization and corruption: evidence across countries”, *Journal of Public Economics* 83 (3):325-345.

Hunt, J. 2007. “How Corruption Hits People When They Are Down”. *Journal of Development Economics*, 84, 574-589.

Molina, E., L. Carella, A. Pacheco, G. Cruces, and L. Gasparini. 2017. "Community monitoring interventions to curb corruption and increase access and quality in service delivery: a systematic review." *Journal of Development Effectiveness* 9 (4):462-499.

Muralidharan, K., & P. Niehaus. 2017. "Experimentation at Scale." *Journal of Economic Perspectives* 31 (4):103-24.

Olken, B. A. 2006. “Corruption and the costs of redistribution: Micro evidence from Indonesia.” *Journal of Public Economics*, 90, 853-870.

Olken, B. 2007. "Monitoring Corruption: Evidence from a Field Experiment in Indonesia." *Journal of Political Economy* 115 (2):200-249.

Olken, B. & Pande, R. 2012. “Corruption in Developing Countries”. *Annual Review of Economics*, 4.

Peters, J., J. Langbein & G. Roberts 2018. Generalization in the Tropics – Development Policy, Randomized Controlled Trials, and External Validity. *The World Bank Research Observer*, 33: 34–64.

Reinikka, R. & Svensson, J. 2004. "Local Capture: Evidence from a Central Government Transfer Program in Uganda. *The Quarterly Journal of Economics*, 119, 679-705.

White, H., Menon, R & H. Waddington, 2018. ”Community-driven development: does it build social cohesion or infrastructure? A mixed-method evidence synthesis.” 3ie Working Paper 30. New Delhi: International Initiative for Impact Evaluation (3ie).

Wong, S. & Guggenheim, S. E.. 2018. ”Community-driven Development : Myths and Realities.” Policy Research Working Paper; no. WPS 8435. Washington, D.C. : World Bank Group.

Woo, J.-Y. 2010. The Impact of Corruption on a Country's FDI Attractiveness: A Panel Data Analysis, 1984-2004. *Journal of International and Area Studies*, 17, 71-91.

Figure 1. Levels of government involved in NUSAF2

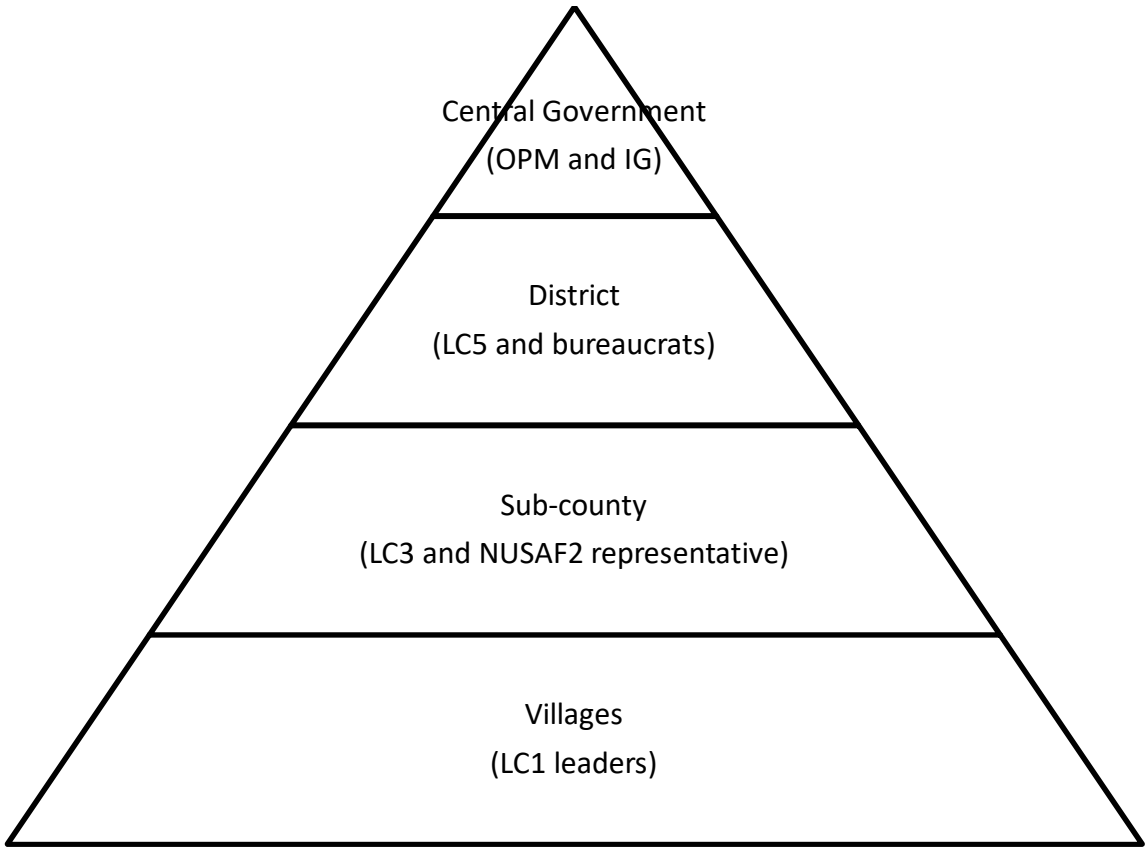


Figure 2. Example of community scorecard

Community Scorecard
NUSAF2 Livestock Sub-Project

Sub-Project Name:

We have compared your sub-project with 33 other sub-projects in your district.

Your Rank

1. Health When Animals Arrived:	21 out of 33
2. Health and Animal Productivity:	22 out of 33
3. Assistance from your DVO:	7 out of 33
4. Value for Money:	20 out of 33

Overall Score: 69 out of 100

Overall Position in the District: **12 of 33**

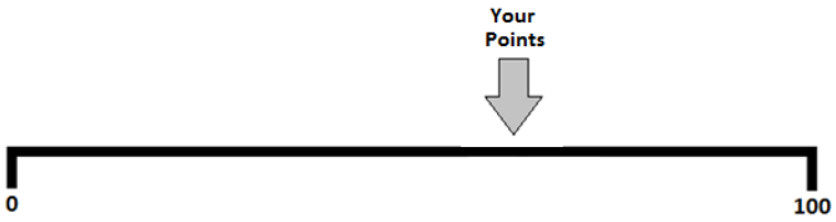


Figure 3. Study Timeline

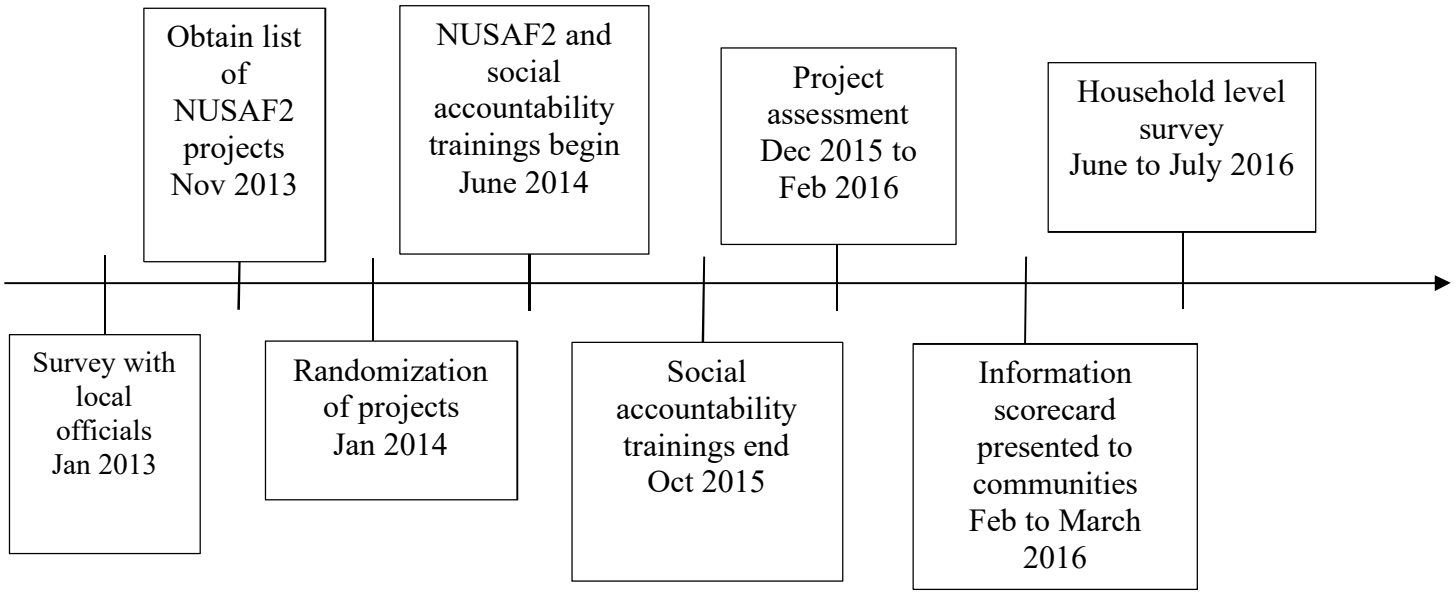


Table 1. Social accountability training design

Project type	Control	Treatment	Total	Control	Treatment	Total	p-value
Livestock	212	423	635	22.6%	45.0%	67.6%	0.904
Enterprise	23	58	81	2.4%	6.2%	8.6%	0.693
Tree Planting	27	47	74	2.9%	5.0%	7.9%	0.177
Staff House	11	36	47	1.2%	3.8%	5.0%	0.763
Road	9	22	31	1.0%	2.3%	3.3%	0.367
Fencing	9	18	27	1.0%	1.9%	2.9%	0.800
Borehole	8	10	18	0.9%	1.1%	1.9%	0.890
OPD	3	8	11	0.3%	0.9%	1.2%	0.623
Dormitory	2	7	9	0.2%	0.7%	1.0%	0.200
Classroom	2	5	7	0.2%	0.5%	0.7%	0.241
Total	306	634	940	32.6%	67.4%	100%	

Notes: This table reports the total number of communities in the social accountability training experiment, as well as their break-down by control and treatment status. The last column provides the p-value for the difference in the share of each project type between treatment and control groups. Due to low numbers of project types, communities below the middle line are not included in the initial analysis (Table 3).

Table 2. Scorecard information design

	Scorecard control	Scorecard treatment	Total
Training control	99	95	194
Training treatment	192	188	380
Total	291	283	574

Notes: This table reports the total number of communities in the scorecard experiment, by scorecard information and social accountability treatment status. As described in the text, to ensure comparability of the projects, the scorecard was designed for the livestock projects only and was implemented everywhere except for the Karamoja region.

Table 3. Balance tests

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Livestock project	Live-stock project is cattle	Project Funds (in 1000s Uganda Schillings)	Project start date (months from Dec 2014)	Project located in sub-county with high reported corruption	Male	Can read (scale 0/2)	Can write (scale 0/2)	Distance to s/c head-quarters
Panel A: All projects									
Training	-0.00302	-0.0132	2,793	-0.00232	0.0269	0.0514***	-0.0118	-0.0935***	-6.233**
	(0.025)	(0.04)	(3905)	(0.003)	(0.034)	(0.02)	(0.02)	(0.03)	(2.90)
Control mean	0.693	0.737864	31,799	0.817	0.316	0.434	.774	.947	79.485
N	940	620	940	940	891	6,218	6,217	6,216	6,142
R-squared	0.65	0	0.62	0.994	0.001	0.075	0.15	0.175	0.249
Panel B: Livestock projects only									
Training only		-0.025	60.24	-0.00644	0.017	0.0215	-0.0116	0.0753	-1.815
		(0.05)	(59.69)	(0.006)	(0.058)	(0.025)	(0.04)	(0.05)	(4.45)
Scorecard only		-0.027	78.43	0.00434	0.043	-0.0305	-0.0326	-0.04	3.665
		(0.06)	(71.07)	(0.007)	(0.067)	(0.028)	(0.05)	(0.05)	(5.36)
Training and scorecard		-0.00022	130.3**	-0.000963	0.0337	0.0164	-0.00176	0.0821*	-1.796
		(0.06)	(59.13)	(0.006)	(0.058)	(0.023)	(0.04)	(0.05)	(4.37)
Control mean		0.747	11,676	1	0.260	0.418	.82	.967	79.592
N		565	574	574	528	3,853	3,851	3,853	3,797
R-squared		0.001	0.764	0.148	0.001	0.075	0.065	0.075	0.194

Notes: This table reports balance tests for project and individual community member characteristics. Panel A contains all projects in the sample. The sample includes 940 projects. Panel B focuses on livestock projects in the scorecard experiment sample. The sample includes 574 projects. Columns 1 to 5 are defined at the project level. Columns 6 to 8 are from the endline household survey and represent participant characteristics that were not expected to change due to the treatment. Standard errors are reported in brackets below the coefficients.
*** p< 0.01, ** p< 0.05, *p< 0.1.

Table 4. Project Score

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Project overall score		Quality score		Quantity score		Project not located	
Training	0.119*	0.123	0.0827	0.104	0.185***	0.167**	-0.00324	-0.00757
	(0.072)	(0.083)	(0.077)	(0.089)	(0.070)	(0.081)	(0.013)	(0.015)
Training*non-livestock		-0.01		-0.0803		0.0724		0.0169
		(0.164)		(0.180)		(0.163)		(0.029)
Non-livestock		0.12		0.262		-0.113		-0.0303
		(0.158)		(0.171)		(0.156)		(0.028)
Control means	-0.024	-0.024	0.005	0.005	-0.066	-0.066	0.027	0.027
N	872	872	871	871	863	863	895	895
R-squared	0.349	0.35	0.35	0.353	0.356	0.356	0.276	0.278

Notes: This table reports the OLS regression results for the treatment effect on project-level outcomes. Odd columns provide the total treatment effect, while even columns include an interaction with whether the project was livestock. The dependent variable in columns 1 and 2 is an aggregate index of columns 3 to 6. Standard errors are reported in brackets below the coefficients. Regressions include region controls. All analysis includes standard errors clustered at the project level.
*** p< 0.01, ** p< 0.05, *p< 0.1.

Table 5. Endline animals and assets

	(1)	(2)	(3)
	Cattle	Livestock index	All assets
Training only (β_1)	0.127	0.079	0.0599
	(0.20)	(0.14)	(0.11)
Scorecard only (β_2)	0.044	0.035	0.0113
	(0.20)	(0.14)	(0.13)
Training and scorecard (β_3)	0.421**	0.303**	0.237**
	(0.19)	(0.14)	(0.12)
Training = Scorecard	0.62	0.72	0.68
Training = Training and scorecard	0.09	0.09	0.09
Scorecard = Training and scorecard	0.04	0.04	0.06
Control mean	2.177	1.915	-0.166
N	3,853	3,853	3,791
R-squared	0.137	0.146	0.137

Notes: This table reports the OLS regression results for the treatment effect on the number of heads of cattle held by the household, a livestock index and total asset index at the endline household survey. The livestock index aggregates different types of livestock using tropical livestock units. Cattle and household asset outcomes were prespecified. Results are provided for the livestock index as an additional robustness check. The livestock index = 0.7 * number of cattle + 0.2 * number of pigs + 0.1 * number of goats and sheep + 0.01 * number of poultry. Variables are top-coded at the 99th percentile. The sample include 3853 beneficiaries in communities from the scorecard experiment sample. Column 1 is reported as number of animals and column 2 is a standardized index of total household assets. Standard errors are reported in brackets below the coefficients. Regressions include subcounty controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table 6. Community monitoring and collective action

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Reporting NUSAF-related issues (total index)	Reporting NUSAF-related issues to LC1	Reporting NUSAF-related issues to Sub-county	Reporting NUSAF-related issues to District	Reporting NUSAF-related issues to IG	Community can solve collective action problems (total index)	Members of the community can come together to solve issues	It is hard for community members to solve issues
Training only (β_1)	0.158*	0.0193	0.0337	0.0254	0.0806***	0.131	0.0943*	-0.0374
	(0.083)	(0.028)	(0.026)	(0.022)	(0.024)	(0.105)	(0.06)	(0.06)
Scorecard only (β_2)	0.220**	0.0569*	0.0318	0.0415*	0.0858***	-0.0509	-0.0276	0.0232
	(0.093)	(0.030)	(0.031)	(0.025)	(0.028)	(0.134)	(0.07)	(0.07)
Training and scorecard (β_3)	0.346***	0.0804***	0.0631**	0.0704***	0.133***	0.192*	0.083	-0.109*
	(0.081)	(0.026)	(0.025)	(0.022)	(0.024)	(0.107)	(0.06)	(0.06)
$\beta_1 = \beta_2$	0.433	0.168	0.946	0.458	0.828	0.111	0.048	0.352
$\beta_1 = \beta_3$	0.006	0.013	0.167	0.013	0.010	0.488	0.812	0.159
$\beta_2 = \beta_3$	0.113	0.376	0.259	0.184	0.061	0.046	0.088	0.052
Control mean	0.837	0.354	0.248	0.147	0.089	0.578	2.999	2.421
N	3,803	3,839	3,841	3,838	3,833	3,848	3,851	3,848
R-squared	0.169	0.129	0.132	0.123	0.144	0.123	0.119	0.117

Notes: This table reports the OLS regression results for the treatment effect on individual-level outcomes at the final endline household survey. Column 1 is an index of columns 2-5, and column 6 is an index of columns 7-8. It is based on the training and scorecard information experiment sample. The sample size is 3853 beneficiaries. Standard errors are reported in brackets below the coefficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table 7. Performance of local leaders

	(1)	(2)	(3)	(4)
	How would you rate performance of sub-project procurement committee?	How would you rate performance of sub-project management committee?	How would you rate performance of social account-ability commit-tee?	How likely would you chose same sub-project man-agement commit-tee?
Training only (β_1)	0.0133	-0.00464	0.00535	-0.0151
	(0.04)	(0.04)	(0.04)	(0.07)
Scorecard only (β_2)	-0.0659	-0.0474	-0.0682*	-0.0225
	(0.04)	(0.04)	(0.04)	(0.08)
Training and scorecard (β_3)	-0.00495	-0.0321	0.0118	-0.0841
	(0.04)	(0.04)	(0.03)	(0.07)
Training = Scorecard	0.0347	0.2591	0.0419	0.901
Training = Training and scorecard	0.5705	0.4001	0.8365	0.1837
Scorecard = Training and scorecard	0.1162	0.6939	0.0357	0.333
Control mean	3.43	3.48	3.42	3.47
N	3,613	3,710	3,588	3,845
R-squared	0.148	0.144	0.128	0.094

Notes: This table reports the OLS regression results for the treatment effect on individual-level outcomes at the final end-line household survey. It is based on the training and scorecard information experiment sample. Leaders' performance is rated on a scale from 1 to 4. The sample size is 3853 beneficiaries. Standard errors are reported in brackets below the co-efficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table 8. Trust in leaders, local officials, and government

	(1)	(2)	(3)	(4)	(5)	(6)
	Project leaders	LC3 chairperson	Sub-county bureaucrats	LC5 chairperson	District bureaucrats	Central government
Training only (β_1)	0.00797	0.0619	0.0796	0.0334	0.0398	0.105***
	(0.044)	(0.059)	(0.049)	(0.063)	(0.058)	(0.037)
Scorecard only (β_2)	-0.00906	0.0233	0.0128	0.0303	-0.0333	0.0936**
	(0.053)	(0.067)	(0.057)	(0.074)	(0.066)	(0.043)
Training and scorecard (β_3)	-0.0555	-0.00962	-0.035	0.00276	-0.0886	0.0701*
	(0.045)	(0.057)	(0.048)	(0.062)	(0.055)	(0.037)
$\beta_1 = \beta_2$	0.687	0.519	0.177	0.964	0.210	0.765
$\beta_1 = \beta_3$	0.064	0.143	0.004	0.577	0.010	0.252
$\beta_2 = \beta_3$	0.301	0.567	0.339	0.678	0.351	0.544
Control mean	3.685	3.204	3.333	3.031	3.332	3.577
N	3,845	3,836	3,822	3,808	3,837	3,831
R-squared	0.119	0.141	0.097	0.083	0.102	0.113

Notes: This table reports the OLS regression results for the treatment effect on individual-level outcomes at the final endline household survey. It is based on the training and scorecard information experiment sample. The sample size is 3853 beneficiaries. Standard errors are reported in brackets below the coefficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level.
*** p< 0.01, ** p< 0.05, *p< 0.1.

Table 9. Heterogeneity in impacts on cattle by level of reported corruption

	(1)	(2)
	High reported corruption	Low reported corruption
Training only (β_1)	0.697	-0.0845
	(0.50)	(0.22)
Scorecard only (β_2)	0.32	0.00235
	(0.47)	(0.22)
Training and scorecard (β_3)	1.413***	0.0833
	(0.52)	(0.19)
Training = Scorecard	0.259	0.669
Training = Training and scorecard	0.062	0.421
Scorecard = Training and scorecard	0.007	0.703
Control mean	2.434	2.113
N	1,013	2,550
R-squared	0.1718	0.1198

Notes: This table reports the OLS regression results for the treatment effect on cattle at the final endline household survey. The cattle variable is top-coded at the 99th percentile. Estimations are based on the training and scorecard information experiment sample (3853 beneficiaries). The sample is split by the whether the sub-county is perceived to have issues of corruption, as reported to the research team during a survey of local officials prior to the roll-out of the interventions. Standard errors are reported in brackets below the coefficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level.
*** p< 0.01, ** p< 0.05, *p< 0.1.

Table 10. Spillovers in community monitoring to other projects

	(1)	(2)	(3)	(4)	(5)
	Reporting other (non-NUSAF) issues (total index)	Reporting other (non-NUSAF) issues to LC1	Reporting other (non-NUSAF) issues to Sub-county	Reporting other (non-NUSAF) issues to District	Reporting other (non-NUSAF) issues to IG
Training only (β_1)	0.0448	0.00219	0.0068	0.0227**	0.0175**
	(0.052)	(0.025)	(0.019)	(0.010)	(0.009)
Scorecard only (β_2)	0.132**	0.0463	0.018	0.0420***	0.0357***
	(0.066)	(0.032)	(0.025)	(0.013)	(0.013)
Training and scorecard (β_3)	0.101*	0.0249	0.0374*	0.0248**	0.0147*
	(0.052)	(0.025)	(0.020)	(0.010)	(0.009)
$\beta_1 = \beta_2$	0.126	0.099	0.587	0.121	0.105
$\beta_1 = \beta_3$	0.202	0.266	0.066	0.829	0.688
$\beta_2 = \beta_3$	0.590	0.417	0.363	0.159	0.067
Control mean	0.384	0.212	0.128	0.032	0.014
N	3,757	3,810	3,810	3,809	3,809
R-squared	0.115	0.13	0.093	0.065	0.057

Notes. This table reports the OLS regression results for the treatment effect on individual-level outcomes at the final endline household survey. It is based on the training and scorecard information experiment sample (3853 beneficiaries). Standard errors are reported in brackets below the coefficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level.
*** p< 0.01, ** p< 0.05, *p< 0.1.

Appendix

A. Curriculum components

The Social Accountability and Community Monitoring training curriculum was developed to be delivered to low-skilled populations, with intensive piloting and heavy focus on visual-based learning. The 7 main modules of the curriculum were as follows.

Module 1: Community Mobilization and Introduction to Social Accountability
This module includes 2 to 3 hours of interaction with mobilized members of the community within which a selected NUSAF2 project is implemented. In the meeting, the community trainer leads a discussion on key concepts of accountability and community engagement, the roles and responsibilities of the Social Accountability Committee (SAC) and conducts the election of 4 willing members of the community to strengthen the existing SAC and form the Community Monitoring Group (CMG).

Part of the discussion includes an overview of NUSAF2 and other existing government programs, and why it is important for the wider community members to monitor these projects even if they are not direct beneficiaries.

Discussions on key concepts of accountability include: a) common types of corruption at the central, local government and community levels such as bribery, embezzlement, nepotism, absenteeism and solicitation of favors; b) social accountability and the constitutional right of every Ugandan to conduct accountability and combat corruption. This session is concluded with brainstorming on key actions the community can take to conduct social accountability, combat corruption and improve project outcomes.

The module ends with the election and introduction of the CMG. Preceding the election, community members are taken through the roles of the monitoring group and characteristics of people who would be suitable for this role. Both the SAC chairman and coordinator of the newly formed CMG are given an opportunity to give short speeches on how they will execute their duties to meet the expectations of the community. The CMG members are then invited for a 3 days training at a selected venue and date.

Module 2: Social Accountability and NUSAF2
A 3-day comprehensive training starts with this module on social accountability and NUSAF2. It reviews into detail all the basic concepts discussed at the mobilization meeting as well as provides a deeper understanding of the different stages of implementation of the NUSAF2 project and the guidelines, for instance the procurement rules and procedures. In this module, the trainer leads the community in identifying key implementation areas that are more prone to mismanagement and explores ways in which the community can engage in monitoring to ensure achievement of the project outcomes.

Module 3: Community Monitoring Skills
This module aims at providing basic skills in community monitoring of NUSAF2 projects. The CMGs are taken through steps in monitoring, identifying sources of information and gathering monitoring data and management of monitoring data. The module includes practical sessions that help CMGs to generate critical questions for monitoring the procurement, timelines, technical

support, financial management and quality of inputs for the NUSAF2 project of their own community.

Module 4: Post-monitoring Activities
This module provides basic understanding on how to review, store and manage monitoring data and outcome. It includes using monitoring data to generate simple monthly reports for submission to relevant authorities. Practical sessions include conducting a mock monitoring session and writing a simple report. The module ends with a session on how to provide feedback on findings from monitoring to the community members as well as explore possible actions to respond to the findings.

Module 5: How to Generate a Community Action Plan
This is a practical step by step session on how to develop an action plan relevant to the project of any given community. CMGs are taken through a participatory discussion that results into key action plans that will be implemented and reviewed with the community trainer during the first follow up support visit. The session includes actual planning and setting timelines for all monitoring activities and allocation of tasks among the CMGs.

Module 6: Follow-up Support Visit
This module provides step by step guidance on how the CMGs can review the action plan generated in module 5 and provide technical support and/or a full refresher training to the CMGs depending on identified technical gaps. The module ends with guidance on how to revise and create new action plans at the end of every follow up support visit.

Module 7: Applying Lessons Learnt to Other Government Services
The aim of this module is to help CMGs apply the monitoring skills they gained from monitoring NUSAF2 to other government programs in their communities. The module uses an example of teacher absenteeism from the education sector to help CMGs learn and apply their skill to other sectors. The module ends with a practical session on creating a monitoring check list using teacher absenteeism as an example, from the original NUSAF2 checklist.

B. Scorecard construction

For the community scorecard, we construct 4 scores for the following dimensions:

- Health when animals arrived
- Animal Productivity
- Assistance from the District Veterinary Officer (DVO)
- Value for money.

We detail how these scores are assigned. All data comes from the project assessment conducted from December 2015 to February 2016.

Health when animals arrived

To construct the health when animals arrived score, we give up to 50 points for the health of animals when they arrived as stated by respondents, and another 50 points for the number of animals that died within 3 months of being received by the respondent.

The 50 points for health as it is stated is constructed by looking at the total number of illnesses identified by respondents within a project (they are asked which illnesses they think each of their animals had when they arrived), divided by the number of animals surveyed. This gives us the average number of illnesses each animal had when they arrived. This is then linearly scaled sending the max average in the dataset to 0, and the minimum average (the fewest illnesses) to 50 points.

For the death within three months, we take the total number of animals that did not die within three months of old age/illness divided by the number of animals they started this. We then multiple this by 50, so that a sub-project gets 50 if no animals died, and 0 if all their animals died of illness/old age.

The final score is then constructed by adding together the respondent health out of 50 and the animal died out of 50, to make a score out of 100.

Animal Productivity

This score is produced by assigning 50 marks for the percentage of animals which are productive (either producing milk or offspring or ploughing) and another 50 marks for the health of the animals measured by the average number of poor health indicators across a variety of indicators. The score is then finally scaled by the number of animals we were able to survey divided by the total number of animals we tried to survey.

For animal productivity we simply define an animal as productive if either it produces milk, has produced calves or is currently able to pull a plough. For example, projects that bought animals that are still too young to be productive get a low score. The score from 50 is the total number of productive animals, divided by the total number of animals we surveyed, and then multiplied by 50.

For the current health of animals, we define a health score for each animal based on the following health indicators: signs of illness, abnormal discharges, skin conditions, parasites, temperament and body score. For each of these indicators each animal gets either a 1 to represent some abnormality or a zero for “healthy”. we then total across all these indicators to give the animal and overall health score (which is an integer between 0 and 6). We then take the mean across the of the animal health scores in the project. Finally, we scale linearly again sending the

project with highest number of illnesses to zero and the project with the lowest average number of illnesses to 50.

To make the final score, we add together the productivity score and the health score. We scaled this score by multiplying by the number of animals were able to survey divided by the number of animals we tried to survey. For example, if in one project we were trying to find 5 cows, we only found 3 but they were perfectly productive and healthy (so would have got a score of 100), then their score will be scaled down to 60 to account for the animals that were not around.

Assistance from the DVO

Assistance from the DVO is constructed using the indicators for the six roles that DVO were supposed to complete for each project. These were: 1) follow-up after inspection, 2) animal treatment/prophylaxis 3) animal ear tagging 4) training project committees 5) animal selection 6) animal inspection.

The first three roles were asked to survey respondents and we assign a score equal to the fraction of respondents that said the DVO provided that service (e.g. 0.6 if 3 of 5 respondents said the DVO ear tagged their animals). The last three roles were asked during the procurement tool to the project committee. For these roles each DVO gets a score of either 0 or 1.

We then sum across these 6 roles, to give a score between 0 and 6. Finally we multiply by 100/6 to give a score from 100.

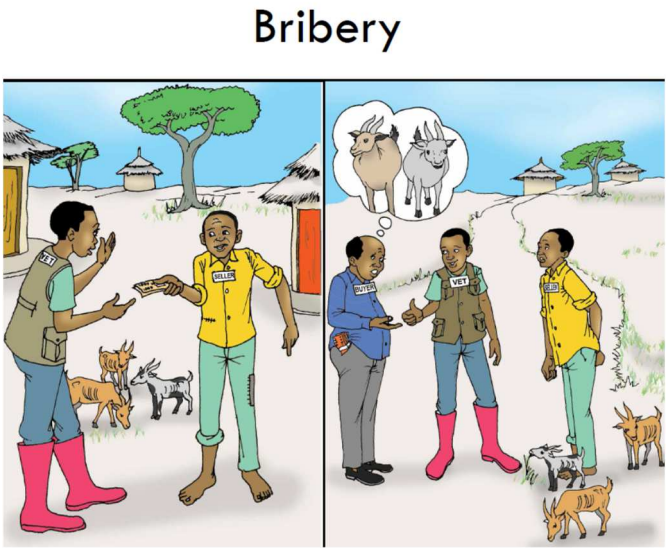
Value for Money

Value for money is constructed using the indicator:

$$\text{VoM} = \frac{Number\ of\ animals\ received \times Producitivity\ score\ of\ all\ animals}{Total\ money\ received\ for\ the\ project}$$

To be able to compare across animal types (cows, goats and sheep), we then adjust this score by standardizing within animal type (subtracting mean and dividing by the standard deviation). Finally, we linearly scale the whole variable, sending the highest deviation above the mean to 100 and the largest deviation below the mean to 0.

Figure A1. Sample of graphics from training



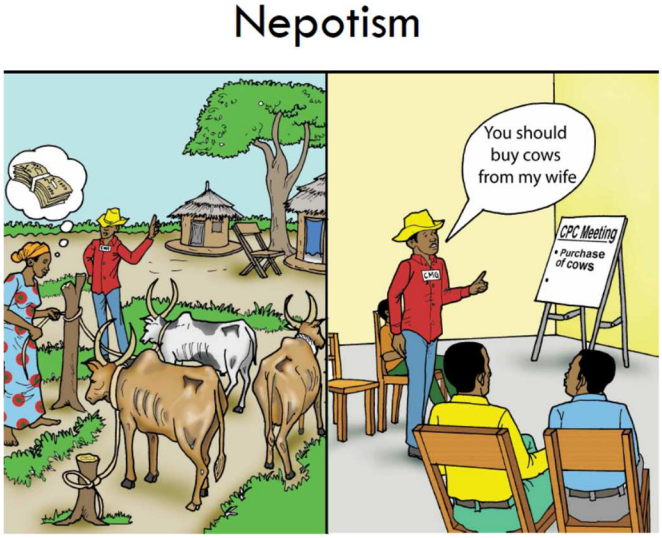
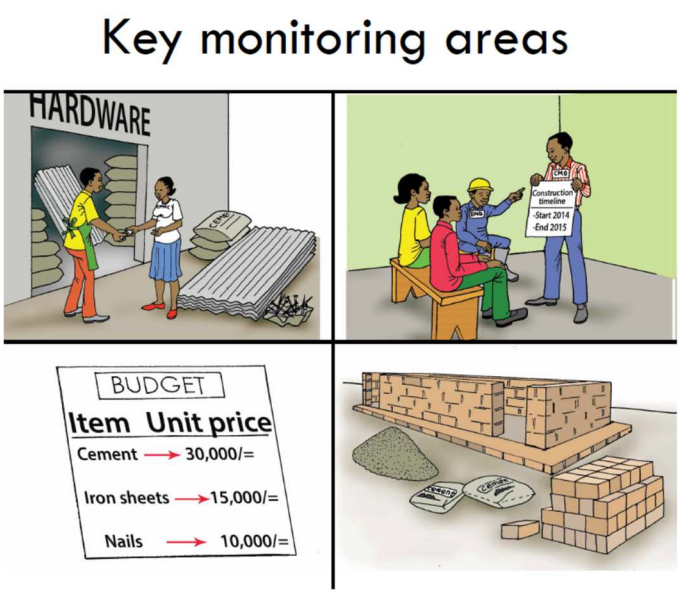
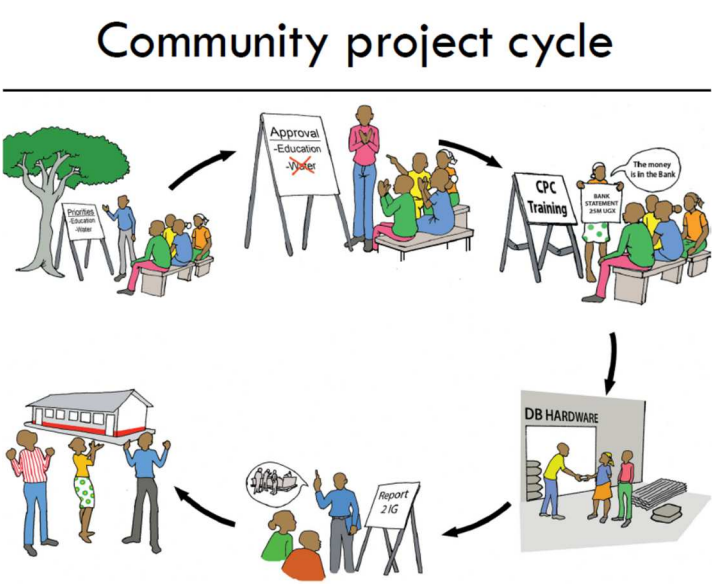


Figure A2. Sample of graphics from training





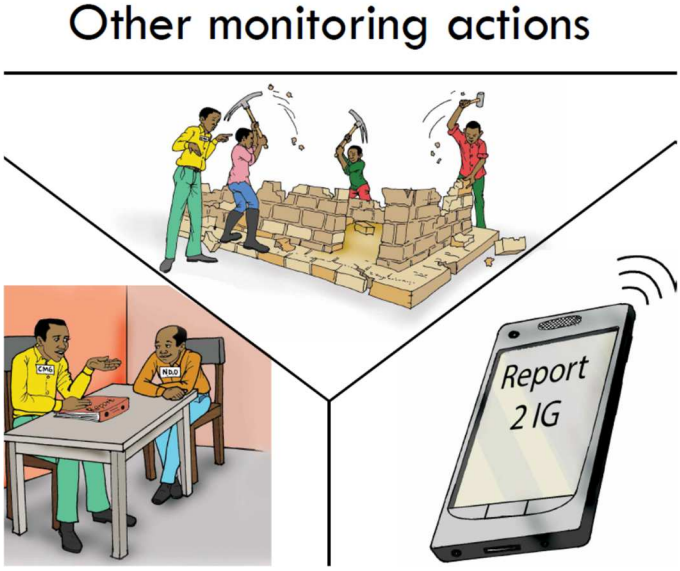


Figure A3. Map of study area

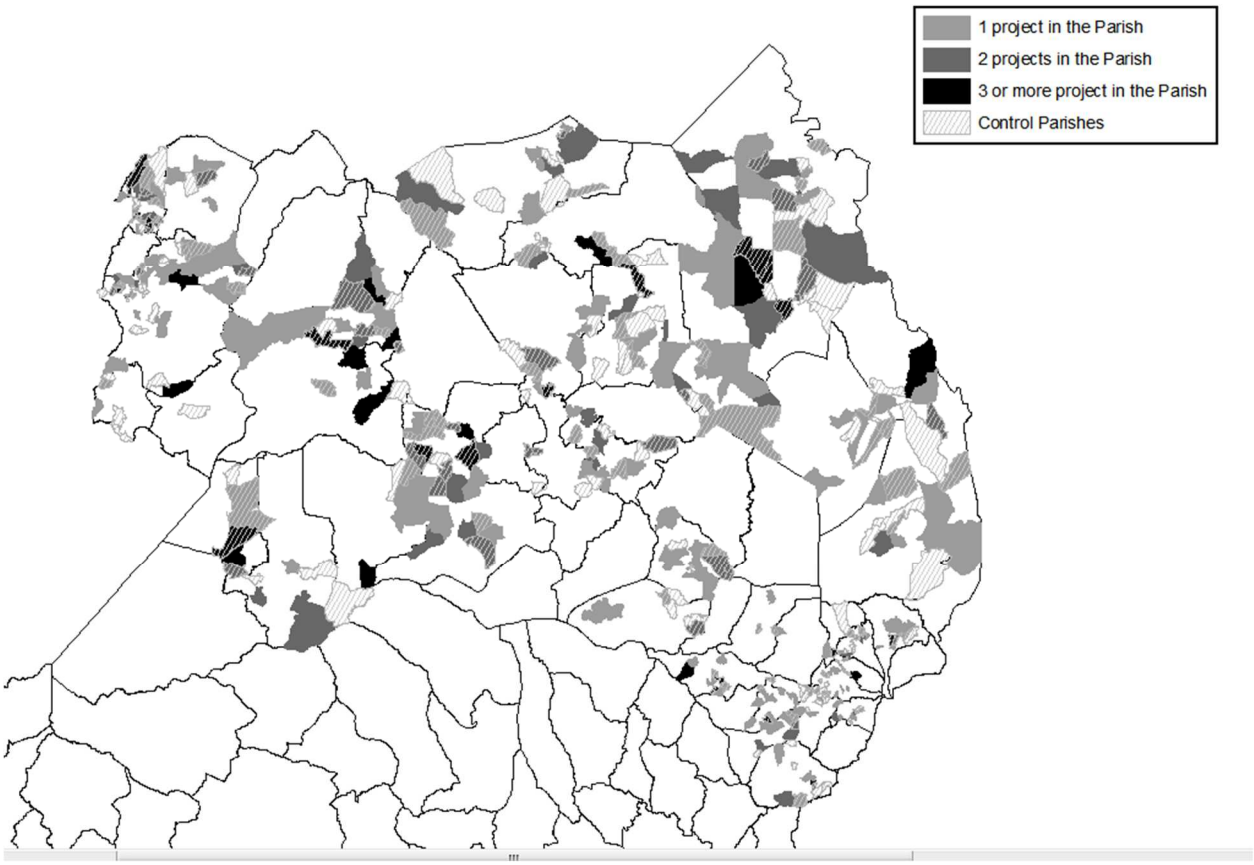


Figure A4. Assessment of a community road project

Photo credit: Mariajose Silva Vargas

Figure A5. Assessment of a livestock project

Photo credit: Mariajose Silva Vargas

Table A1. Project score construction

Subproject Type	Quantity Score		Quality Score		
	Unit	Score	Quality Indicators	Construction	Score
Livestock	Animals	Total number of animals received	1. Correct age of the animal when it was received	Binary indicator for correct age of animal, i.e. 2 year to 4 years for male cows and 2.5 - 4.5 years for female cows	Average of Quality Indicators
			2. Improved breed of the animal	Binary indicator which takes 1 if the animal received is improved breed	
			3. Productivity of the animal	Binary indicator which takes 1 if the animal did at least one of the followings: oxen ploughing, given birth (female), bull breeding, pregnant (female cows and goats/sheep), giving milk and female cow ploughing	
			4. Animal health	Binary illness indicator which takes 1 if the animal has at least one illness. Note: 50% of the animals observed did not have any illness	
Staff House	M^2	Size of the staff house built	1. Walls	Binary indicator which takes 1 if the part is completed to a satisifactory standard	Average of Quality Indicators
			2. Roof		
			3. Ceiling		
			4. Floor		
			5. Painting		
			6. Doors	Binary indicator which takes 1 if there is at least one is built and functioning	
			7. Windows		
			8. Electricity	Binary indicator for having power supply that is complete	
			9. Water Tank	Binary indicator for having water tank built	
Enterprise	People	The number of people currently invloved in the enterprise	1. Equipment	Binary indicator for having secure access to each category for business	Average of Quality Indicators
			2. Materials		
			3. Transportation		
			4. Credit		
			5. Skilled labour		
			6. Markets		
			6. Markets		
			7. Success	Binary indicator which takes 1 if the enterprise owner feels the business is successful	
Fencing	M	Length of the fence	1. Fence	Binary indicator for completion of each category	Average of Quality Indicators
			2. Main gate		
			3. Small gate		
			4. Guard house		
Roads	M^2	Road surface area	1. Material of the road	Binary indicator for gravel road (entirely or mixed as opposed to earth/dirt)	Average of Quality Indicators
			2. Road surface	Binary indicator for satsifactory road surface	
			3. Wingwalls	Binary indicator for at least one satisfactory wingwall but none defective	
			4. Drainage lines	Binary indicator for satsifactory status of each category	
			5. Scour checks		
			6. Mitre drains		
			7. Culverts		
Tree Planting	Acres	Total amount of land in acres	1. Seed certification	Binary indicator which takes 1 if the batch of seeds/seedlings came with a certification number	Average of Quality Indicators
			2. Herbicide	Binary indicator for having sprayed with herbicides during pre-planting	
			3. Training	Average of 7 binary indicators for having received advice on (1) species selection, (2) weeding, (3) planting preparation, (4) disease detection and treatment, (5) fire prevention, (6) pruning/thinningm and (7) record keeping	

Table A2. Other index construction

Category	Index	Range	Description	Variables
Implementation	**Project Implementation Quality Index**	0 - 4	Additive index with sum of 4 discrete variables, each of which describes how the project implementation was perceived by beneficiaries	1. Project usefulness (0-1)
				2. Project completed (0/1)
				3. Satisfaction with material (0/1)
				4. Satisfaction with cost of material (0/1)
Procurement	**Challenges in Procurement Process Index**	0 - 4	Additive index with sum of 4 binary variables, each of which indicates challenges/violations in procurement process	1. Funds withdrawn by members outside of CPMC (0/1)
				2. Project material acquired by members outside of CPC (0/1)
				3. Less than three steps taken to purchase materials (0/1)
				4. Procurement process was difficult (0/1)
	Satisfaction with supplier Index	0 - 8	Additive index with sum of 2 discrete variables	1. Relationship with the local suppliers (0-4)
				2. Level of satisfaction with the services provided by the supplier (0-4)
	Hired a Contractor to Implement Project	0 / 1	Binary indicator for hiring a contractor	1. Hired a Contractor to Implement Project (0/1)
	Index of challenges in Contracting Process	0 - 9	Additive index with sum of 9 binary variables, each of which indicates challenges/violations in procurement process conditional on hiring a contractor	1. No advertisement to select contractor (0/1)
				2. There were less than 3 bidders (0/1)
				3. Bids not registered (0/1)
				4. Less than 2 (out of 5 advised) contacting steps involved (0/1)
				5. No information gathered on contractor during vetting process (0/1)
				6. Oustide influence in the contractor selection process (0/1)
				7. Contractor not signed a formal contract (0/1)
				8. Beneficiary not consulted during implementation (0/1)
				9. Beneficiary contribution not taken into consideration (0/1)
	Satisfaction with contractor Index	0 - 8	Additive index with sum of 2 discrete variables	1. Relationship with the contractor/local lead artisan (0-4)
				2. Level of satisfaction with the services provided by the contractor (0-4)
Monitoring	**Index for Intensity of Project Community Monitoring**	0 - 4	Additive index with sum of 4 binary variables	1. Compiled an Accountability Report (0/1)
				2. Monitored project implementation (0/1)
				3. Monitored selection of materials/livestock (0/1)
				4. Monitoring report was written (0/1)
	Index for Intensity of Social Accountability Committee Project Monitoring	0 - 4	Additive index with sum of 4 binary variables, each of which indicates SAC involvement and quality	1. SAC witnessed delivery of procured goods (0/1)
				2. SAC wrote monitoring report (0/1)
				3. SAC monitored project implementation (0/1)
				4. SAC monitored selection of materials/livestock (0/1)
Interactions with Leaders	**Satisfaction with NUSAF Desk Officer (NDO) Index**	0 - 8	Additive index with sum of 2 discrete variables	1. Relationship with the NDO (0-4)
				2. Level of satisfaction with the services provided by the NDO (0-4)
	Satisfaction with Disctrict Vet Officer (DVO) Index	0 - 8	Additive index with sum of 2 discrete variables	1. Relationship with the DVO (0-4)
				2. Level of satisfaction with the services provided by the DVO (0-4)
Reporting	**Reporting NUSAF-Related Issues**	0 - 2	Additive index with sum of 2 binary variables	1. Beneficiary reported NUSAF-related issues (0/1)
				2. Someone else in the group reported NUSAF-related issues (0/1)
Trust	**Trust**	1 - 4	Single categorical variable	1. Level of trust in leaders (1-4)

Table A3. Descriptive Statistics

	(1) Mean	(2) SD	(3) Min	(4) Max	(5) Obs
Project level:					
Project Funds (in 1,000 ugx)	22750.6	32741.9	7612	162670	895
Livestock project (0/1)	0.709	0.454	0	1	895
Project start date (Period when grants were received)	38.188	3.935	1	48	812
Project overall score (std)	0.032	0.949	-2.9	3.28	872
Project quality score (std)	0.043	1.021	-2.86	3.23	871
Project quantity score (std)	0.016	0.928	-5.71	12.1	863
Project is missing (0/1)	0.027	0.162	0	1	895
Project Implementation Quality Index	2.38	0.905	0	4	705
Satisfaction with supplier Index	5.629	1.366	0	8	660
Hired a Contractor to Implement Project	0.388	0.487	0	1	800
Index of challenges in Contracting Process	3.332	1.868	0	9	301
Satisfaction with contractor Index	5.375	1.77	0	8	459
Index for Intensity of Project Community Monitoring	2.957	1.139	0	4	821
Index for Intensity of Social Accountability Committee Project Monitoring	0.831	0.954	0	4	863
Satisfaction with NDO Index	5.912	1.407	0	8	839
Satisfaction with District Vet Index	10.638	1.463	6	15	572
Animal level:					
Animal dead (0/1)	0.13	0.336	0	1	6891
Animal sold (0/1)	0.051	0.22	0	1	6891
Animal stolen (0/1)	0.017	0.131	0	1	6891
Animal dead/sold/stolen (0/1)	0.198	0.399	0	1	6891
Beneficiary level:					
Number of Cattle (Total)	2.452	10.5	0	800	6961
Number of Goats (Total)	4.206	7.018	0	230	6966
Number of Livestock in Tropical Livestock Unit	1.816	5.523	0	406.5	6952
Reporting NUSAF-related issues (total)	1.052	1.310	0	4	6966
Reporting NUSAF-related issues to LC1	0.405	0.491	0	1	6964
Reporting NUSAF-related issues to Subcounty	0.305	0.460	0	1	6961
Reporting NUSAF-related issues to District	0.203	0.402	0	1	6963
Reporting NUSAF-related issues to IG	0.141	0.348	0	1	6957
Trust Project Leaders (1-4)	3.582	0.712	1	4	6952
Trust LC3 Chairperson (1-4)	3.181	0.93	1	4	6937
Trust Subcounty Bureaucrats (1-4)	3.295	0.809	1	4	6921
Trust LC5 Chairperson (1-4)	2.997	1.011	1	4	6892
Trust District Bureaucrats (1-4)	3.297	0.859	1	4	6933
Trust Government (1-4)	3.635	0.66	1	4	6932

Table A4. Project Scores, robustness to top-coding

	(1)	(2)	(3)
	Project overall score (topcoded)	Quality score (topcoded)	Quantity score (topcoded)
Training	0.118*	0.085	0.109**
	(0.07)	(0.08)	(0.05)
Control means	-0.024	0.002	-0.061
N	872	871	863
R-squared	0.349	0.353	0.415

Notes: This table reports the OLS regression results for the treatment effect on project-level outcomes. The dependent variable in columns 1 is an aggregate index of columns 2 and 3. Variables are top-coded above the 99th percentile. Standard errors are reported in brackets below the coefficients. Regressions include region controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table A5. Livestock project score components (at beneficiary level)

	(1)	(2)	(3)	(4)	(5)
	Fraction of Animals dead, stolen or sold	Animal Bought at the Correct Age (0/1)	Animal Is an Improved/Crossed/Hybrid Breed (0/1)	Animal Is Productive (0/1)	Animal Health by Mean Number of Illnesses
Training (γ_1)	-0.0463***	0.0229	0.0156	-0.0127	0.0276*
	(0.02)	(0.03)	(0.01)	(0.03)	(0.01)
Control mean	0.187	0.346	0.167	0.397	0.860
N	3,044	3,044	3,040	2,850	2,850
R-squared	0.171	0.282	0.798	0.222	0.175

Notes: This table reports the OLS regression results for the treatment effect on components of the livestock project score (from the sub-project assessment), aggregated across animal at the beneficiary level. The illness index is re-weighted as 1 minus the mean number of illnesses, so the positive coefficient means fewer observed illnesses. Standard errors are reported in brackets below the coefficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table A6. Community Monitoring (all projects)

	(1)	(2)
	Index for Intensity of Project Monitoring by Community	Index for Intensity of Project Monitoring by Social Accountability Committee
Training	0.155*	0.248***
	(0.08)	(0.07)
Control means	2.767	0.628
N	855	907
R-squared	0.477	0.427

Notes: This table reports the OLS regression results for the training treatment effect on project-level intermediary outcomes, including all projects in the training experiment sample. Intermediary outcomes include indicators of community monitoring, such as an index of the intensity of project community monitoring, and an index of the intensity of social accountability committee (SAC) project monitoring. Standard errors are reported in brackets below the coefficients. Regressions include region controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table A7. Community monitoring (all projects)

	(1)	(2)	(3)	(4)	(5)
	Reporting NUSAF-related issues (total index)	Reporting NUSAF-related issues to LC1	Reporting NUSAF-related issues to Sub-county	Reporting NUSAF-related issues to District	Reporting NUSAF-related issues to IG
Training only	0.164***	0.0335**	0.0327**	0.0329**	0.0654***
	(0.05)	(0.02)	(0.02)	(0.01)	(0.01)
Control mean	0.959	0.386	0.288	0.186	0.102
N	6,125	6,198	6,190	6,190	6,186
R-squared	0.158	0.127	0.126	0.115	0.134

Notes: This table reports the OLS regression results for the training treatment effect on individual-level outcomes at the endline household survey, including all projects in the training experiment sample. Standard errors are reported in brackets below the coefficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table A8. Impacts on procurement of project output

	(1)	(2)	(3)	(4)	(5)
	Challenges in procurement process index	Satisfaction with supplier index	Hired a contractor (0/1)	Index of challenges in contracting process	Satisfaction with contractor index
Training	-0.0566	0.118	0.001	-0.182	-0.0411
	(0.08)	(0.13)	[0.974]	(0.25)	(0.20)
Control means	2.046	5.544	0.406	3.623	5.273
N	742	686	800	307	501
R-squared	0.272	0.336	0.615	0.532	0.391

Notes: This table reports the OLS regression results for the training treatment effect on project-level intermediary outcomes, including all projects in the training experiment sample. Intermediary outcomes include an index of challenges faced by communities in the procurement process, an index of satisfaction with suppliers of goods and materials, and whether the community hired a contractor. Some variables are coded as missing if communities did not procure goods or services (column 1), did not use suppliers (column 2), or did not use contractors (column 4 and 5). Standard errors are reported in brackets below the coefficients. Regressions include region controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table A9. Interactions with bureaucrats (all projects)

	(1)	(2)	(3)	(4)
	Payment was made to district official	Payment was made to district officer (vet, engineer, etc.)	Satisfaction with NDO Index	Satisfaction with District Vet Index
Training	-0.00361	-0.0143	-0.0187	0.0933
	(0.03)	(0.03)	(0.11)	(0.13)
Control means	0.141	0.166	5.822	10.566
N	871	871	881	572
R-squared	0.345	0.317	0.301	0.429

Notes: This table reports the OLS regression results for the training treatment effect on project-level intermediary outcomes, including all projects in the training experiment sample. Intermediary outcomes include indicators on whether a payment was made to a district official or staff, and an index of satisfaction with the sub-county NUSAF2 official and district veterinarian officer. Standard errors are reported in brackets below the coefficients. Regressions include region controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table A10. Performance of local leaders (all projects)

	(1)	(2)	(3)	(4)
	How would you rate performance of sub-project procurement committee overall	How would you rate performance of sub-project management committee overall	How would you rate performance of social accountability committee overall	How likely would you chose same sub-project management committee
Training only (β_1)	0.0267	-0.00357	0.0313	-0.0383
	(0.02)	(0.02)	(0.02)	(0.04)
Control mean	3.36	3.42	3.35	3.38
N	5,773	5,940	5,797	6,205
R-squared	0.134	0.128	0.115	0.104

Notes: This table reports the OLS regression results for the training treatment effect on individual-level outcomes at the final endline household survey, including all projects in the training experiment sample. Leaders' performance is rated on a scale from 1 to 4. Standard errors are reported in brackets below the coefficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table A11. Trust in leaders, local officials, and government (all projects)

	(1)	(2)	(3)	(4)	(5)	(6)
	Project leaders	LC3 chairperson	Sub-county bureau-crats	LC5 chairperson	District bureau-crats	Central government
Training only	-0.0628**	-0.021	0.00296	-0.0141	-0.0159	0.0386*
	(0.03)	(0.03)	(0.03)	(0.04)	(0.03)	(0.02)
Control mean	3.652	3.197	3.309	3.022	3.324	3.620
N	6,205	6,192	6,175	6,155	6,190	6,187
R-squared	0.115	0.135	0.115	0.103	0.104	0.109

Notes: This table reports the OLS regression results for the training treatment effect on individual-level outcomes at the endline household survey, including all projects in the training experiment sample. Trust is rated on a scale from 1 to 4. Standard errors are reported in brackets below the coefficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table A12. Spillovers of reporting issues to officials within communities (all projects)

	(1)	(2)	(3)	(4)	(5)
	Reporting other (non-NUSAF) issues (total index)	Reporting other (non-NUSAF) issues to LC1	Reporting other (non-NUSAF) issues to Sub-county	Reporting other (non-NUSAF) issues to District	Reporting other (non-NUSAF) issues to IG
Training only	0.0895***	0.0244	0.0370***	0.0202***	0.00635
	(0.03)	(0.01)	(0.01)	(0.01)	(0.01)
Control mean	0.441	0.228	0.132	0.057	0.027
N	6,154	6,152	6,152	6,147	6,147
R-squared	0.125	0.095	0.078	0.058	0.058

Notes. This table reports the OLS regression results for the training treatment effect on individual-level outcomes at the endline household survey, including all projects in the training experiment sample. Standard errors are reported in brackets below the coefficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table A13. Spillovers across communities

Training x scorecard	0.525**
	[0.226]
Intensity of treatment	23.241
	[51.952]
Control means	2.207
N	3,851
R-squared	0.085

Notes: This table reports the OLS regression results for the treatment effect on household-level cattle outcome at the endline survey. Standard errors are reported in brackets below the coefficients. Regressions include region controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table A14. Additional heterogeneities in impacts on cattle at endline (trimmed variable)

	(1)	(2)	(3)	(4)	(5)	(6)
Training only	0.215	0.0893	-0.0261	0.11	0.0597	0.0391
	(0.22)	(0.23)	(0.27)	(0.21)	(0.23)	(0.41)
Scorecard only	0.118	0.333	-0.0349	0.0955	-0.0403	-0.0944
	(0.20)	(0.24)	(0.26)	(0.21)	(0.23)	(0.39)
Training and scorecard	0.547***	0.725***	0.191	0.379*	0.349	0.145
	(0.21)	(0.25)	(0.27)	(0.20)	(0.22)	(0.33)
Training only *male	-0.233					
	(0.27)					
Scorecard only * male	-0.138					
	(0.29)					
Scorecard and Training * male	-0.314					
	(0.28)					
male	0.676***					
	(0.20)					
Training only * read		0.0269				
		(0.27)				
Scorecard only * read		-0.482*				
		(0.29)				
Scorecard and Training * read		-0.512*				
		(0.28)				
Read		0.681***				
		(0.21)				
Training only * distance			0.00185			
			(0.00)			
Scorecard only * distance			0.00123			
			(0.00)			
Scorecard and Training * distance			0.00245			
			(0.00)			
Distance			-0.00616***			
			(0.00)			
Training only * connected to lc1				0.0764		
				(0.39)		
Scorecard only * connected to lc1				-0.377		
				(0.40)		
Scorecard and Training * connected to lc1				0.0696		
				(0.44)		
Connected to lc1				0.367		
				(0.30)		
Training only * regular beneficiary					0.0462	
					(0.26)	
Scorecard only * regular beneficiary					0.21	
					(0.27)	
Scorecard and Training * regular beneficiary					0.0453	
					(0.28)	
Regular beneficiary					-0.518**	
					(0.21)	
Training only * cattle project						0.0547
						(0.46)
Scorecard only * cattle project						0.196
						(0.44)
Scorecard and Training * cattle project						0.357
						(0.40)
cattle project						1.323***
						(0.39)
Control mean	2.17	2.17	2.17	2.17	2.17	2.17
N	3,853	3,853	3,797	3,758	3,849	3,791
R-squared	0.144	0.144	0.145	0.135	0.142	0.161

Notes: This table reports the OLS regression results for the treatment effect on individual-level outcomes at the endline household survey. It is based on the training and scorecard information experiment sample (3853 beneficiaries). Standard errors are reported in brackets below the coefficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.

Table A15. Endline animals and assets robustness test

	(1)	(2)	(3)
	Cattle	Livestock index	All assets
Training only (β_1)	0.162	0.116	0.145
	(0.194)	(0.153)	(0.142)
Scorecard only (β_2)	0.0084	0.0243	0.0472
	(0.208)	(0.162)	(0.165)
Training and scorecard (β_3)	0.508**	0.366**	0.309**
	(0.242)	(0.186)	(0.148)
Control mean	2.177	1.915	-0.166
N	3,686	3,680	3,628
R-squared	0.137	0.146	0.137

Notes: This table reports the OLS regression results for the treatment effect on the number of heads of cattle held by the household, a livestock index and total asset index at the endline household survey. Two regular members were randomly dropped as a robustness test. The livestock index aggregates different types of livestock using tropical livestock units. Cattle and household asset outcomes were pre-specified. Results are provided for the livestock index as an additional robustness check. The livestock index = 0.7 * number of cattle + 0.2 * number of pigs + 0.1 * number of goats and sheep + 0.01 * number of poultry. Variables are top-coded at the 99th percentile. The sample include 3853 beneficiaries in communities from the scorecard experiment sample. Column 1 is reported as number of animals and column 2 is a standardized index of total household assets. Standard errors are reported in brackets below the coefficients. Regressions include sub-county controls. All analysis includes standard errors clustered at the project level. *** p< 0.01, ** p< 0.05, *p< 0.1.