Policy Research Working Paper 9875

Building State Capacity

What Is the Impact of Development Projects?

Vincenzo Di Maro
David K. Evans
Stuti Khemani
Thiago De Gouvea Scot de Arruda

WORLD BANK GROUP
Development Economics
Development Impact Evaluation Group
December 2021

Policy Research Working Paper 9875

Abstract

Although research has established the importance of state capacity in economic development, less is known about how to build that capacity and the role of external partners in the process. This paper estimates the impact of a typical development project designed to build state capacity in a low-income country. Specifically, it evaluates a multilateral development bank project in Tanzania, which incentivized investments in local state capacity by offering grants conditional on institutional performance scores. The paper uses a difference-in-differences methodology to estimate the project impact, comparing outcomes between 18 project and 22 non-project local governments over 2016–18. Outcomes were measured through two rounds of primary surveys of nearly 500 local government officials and nearly 3,000 households. Over the course of the project, measured state capacity improved in project areas, but due to comparable gains in non-project areas, the project's value-added to change in state capacity is estimated to be zero across all the dozens of relevant variables in the surveys. The data suggest that state capacity is evolving in Tanzania through endogenous changes in trust and legitimacy in the country rather than from financial incentives offered by external partners.

This paper is a product of the Development Impact Evaluation Group, Development Economics. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at vdimaro@worldbank.org, devans@cgdev.org, skhemani@worldbank.org, and tscot@worldbank.org.

Building State Capacity: What Is the Impact of Development Projects?*

Vincenzo Di Maro David K. Evans Stuti Khemani Thiago Scot

Originally published in the Policy Research Working Paper Series on *December 2021.* This version is updated on *February 2022.*
To obtain the originally published version, please email prwp@worldbank.org.

Keywords: state capacity, governance, decentralizations, performance-based financing
JEL Codes: D02, O10, O19

*For support in implementing this evaluation, we thank Gilbert Mfinanga, Lucy Mzengi, Davis Shemangale, and others in the Tanzania's President's Office, Regional Administration and Local Government Tanzania (PO-RALG). We thank Arianna Legovini and Aart Kraay for guidance and feedback throughout this work, and Macartan Humphreys, Philip Keefer, Thomas Kenyon, Bob Rijkers, three anonymous reviewers, and seminar participants at the WZB Berlin Social Science Center for comments. Anna Popova and Susana Cordeiro Guerra provided excellent research assistance for this project. Funding for this impact evaluation was provided by the Impact Evaluation for Development Impact (i2i) program. The authors' affiliations and contact are Di Maro (World Bank: vdimaro@worldbank.org), Evans (Center for Global Development: devans@cgdev.org), Khemani (World Bank: skhemani@worldbank.org), and Scot (World Bank: tscot@worldbank.org).

"Little else is requisite to carry a state to the highest degree of opulence from the lowest barbarism, but peace, easy taxes, and a tolerable administration of justice; all the rest being brought about by the natural course of things." –attributed to Adam Smith in Stewart (1795)

1 Introduction

State capacity is essential for economic development.[1] The hallmark of countries characterized as "developing," which distinguishes them from "developed" countries, is the lack of some basic features of state capacity that prosperous societies possess (Besley & Persson, 2011; Herbst, 2000; Migdal, 1988). For example, countries with limited state capacity may lack the ability to raise taxes to finance basic public services, enforce contracts, or protect property rights. International development agencies have thus invested in capacity building projects to strengthen state institutions in developing countries. Although research has established the importance of state capacity in economic development, little is known about how state capacity is built and what is the role of external development partners in this regard (Besley & Persson, 2009; Acemoglu et al., 2015; Fergusson et al., 2020).[2] This paper addresses that gap, providing quantitative results from the evaluation of a typical World Bank project designed to strengthen the capacity of the local state.

Since the 1990s, the World Bank has invested in various projects across regions to strengthen the capacity of the local state (Independent Evaluation Group, 2008, 2018). The core design features are shared by multiple World Bank projects which aim to strengthen local governments across regions. Hence, while this study reports the impacts of such a project in Tanzania, the findings may have implications for how a major global development agency approaches state capacity building projects around the world.

The essential design feature is to provide incentives to local governments to undertake verifiable actions that the project regards as essential for state capacity. These actions include showing

[1]This statement is both axiomatic in economics as well as one that has recently been supported by rigorous evidence. Modern economics starts with the assumption that state capacity exists to protect property rights, maintain law and order, enforce contracts, and collect the tax revenues needed for these public services (Dincecco & Katz, 2016; Acemoglu, García-Jimeno, & Robinson, 2015; Besley & Persson, 2011) One strand of empirical work finds a robust positive correlation between tax to GDP ratios, and other measures of state capacity, with economic development (Acemoglu, 2005; Besley & Persson, 2009; Dincecco & Katz, 2016). Another strand finds persistent differences in income and prosperity across places which have a history of strong state institutions versus those that do not (Bandyopadhyay & Green, 2016; Gennaioli & Rainer, 2007; Michalopoulos & Papaioannou, 2013).

[2]Qualitative discussion of whether World Bank projects build state capacity is provided in de Janvry & Dethier (2012) and in Stern, Dethier, & Rogers (2005). Some previous studies have examined the cross-country relationship between World Bank projects and subsequent state capacity, finding a suggestive positive relationship for highly rated projects (Hanson & Sigman, 2019). A recent paper, Erman et al. (2021), uses a fixed effects identification strategy and administrative data on municipal revenues to examine the impact of a project in Mozambique. They find that project municipalities increased own revenue collection over time more than non-project municipalities. In contrast, as we report below, not only do we find no impact of the project in Tanzania on local tax payment and attitudes, but also a striking illustration of how context matters for project design. While the project in Tanzania had initially targeted own-source revenues as a key component of the performance grant, it dropped this component after was re-centralized . Nevertheless, we find evidence of increasing willingness to pay property taxes–a measure of state capacity–in both project and non-project urban local governments in Tanzania.

records of urban planning, internal audits, establishment of tender boards for public procurement, recruitment of key state personnel, adoption of technologies for property tax management, and increase in local own-source revenues, among others. The size of fiscal grants to local governments under the project are conditional on their performance in achieving institutional outcomes, as assessed by an independent third-party. On average, the fiscal grants represent about 15 percent of total revenues for local governments in 2011.[3] In addition to the performance grants, these projects also include mechanisms to help local governments with advice and technical assistance on how to plan, audit, procure, recruit personnel, use technology, increase revenue collection, etc. In other words, local governments receive technical support in all the areas of institutional performance on which they are being assessed.

This paper provides evidence on whether institutional outcomes targeted by the capacity building project in Tanzania are different across local governments that received the project and those that did not. Since the selection of project local governments is not amenable to randomization in these types of programs,[4] any evaluation design has to contend with both observable and unobservable initial differences between treatment (those receiving the World Bank capacity building project) and comparison (those not receiving the project) Local Government Areas (LGAs). To address these issues we use a difference-in-differences method to estimate the project impact by comparing outcomes between project and non-project LGAs over time. The difference-in-differences method does not require treatment and comparison areas to be statistically identical at the start of the program, but it does require that the two groups would develop at similar rates in the absence of the project. This is often called the "parallel trends" assumption.

The evidence comes from 40 Local Government Areas of Tanzania over the period 2016-2018, of which 18 LGAs had been selected in 2013 to receive the World Bank capacity building project. The project team indicated at the outset that the selection of these 18 LGAs was purposive, based on negotiation between the World Bank and the Government of Tanzania (World Bank, 2012). Thus, we knew at the outset that these 18 LGAs were likely to be systematically different from other LGAs in the country on both observable and unobservable variables. To minimize these differences, we worked closely with the government and the World Bank project team to identify comparable LGAs (22 non-project LGAs were identified to serve as comparators) which would not be receiving the project, complementing this identification process with propensity score matching. The difference-in-difference strategy we adopted was supported at the outset by evidence that in the years preceding the project, project and non-project LGAs had similar growth in local, "own source" revenues, a key measure of state capacity. This similar growth in the absence of the intervention is the fundamental assumption behind the difference-in-differences estimation strategy.

We then implemented two rounds of surveys across these 40 LGAs (18 receiving the project and 22 comparators) of 474 government officials and 2,998 citizens, two years apart, during a

[3]Total revenue includes own source revenues and transfers from the central government. 2011 is the most recent year for which we have revenue data disaggregated at the LGA level.

[4]We discuss the infeasibility of randomization in the section describing the institutional context.

period in which the project was well established, had disbursed and was expected to have undertaken capacity building activities. The first survey was undertaken in February 2016, and the second in April 2018. The surveys included a variety of questions about the capacity of local government officials to plan, recruit personnel, manage public funds, raise revenues and deliver services, and the experience of citizens in receiving these services.

Across all the survey measures, we find no evidence of significant differences between the 18 project and 22 non-project LGAs over the two years during which the project disbursed its funds and enabled technical assistance to build local state capacity. Specifically, many indicators improved in project LGAs, but they improved at similar rates in non-project LGAs. Some of the estimates are precisely estimated at 0, based on the responses of government officials. For example, on questions of whether master urban plans have been updated (and can be verified by the interviewer), whether internal audits have been undertaken (and can be verified by the interviewer), whether budgets are executed in a timelier fashion, and efforts towards own revenue collection, there is similar improvement over time in both project and non-project LGAs. That is, non-project LGAs were able to make comparable improvements even in the absence of the incentives and training available under the project.

We discuss alternative explanations for our results, beyond the possibility of limited project impacts to the ongoing evolution of state capacity in Tanzania. Perhaps the two most important of these alternative possibilities are the following. First, the project's effects might have been concentrated at the early stages of implementation, before the first survey we undertook in February 2016. Second, the project LGAs' performance might have played the role of demonstrating to non-project LGAs the importance of investing in their own capacity or, more generally, copying project LGAs' practices. Addressing the first point, we show that that the project's own scoring of local government performance does not indicate that all improvement across the project life-cycle happened before our baseline survey. Furthermore, approximately half of the total disbursement of the project happened between baseline (or first round of survey) and endline (or second round), so the period between our surveys covers a substantial amount of the project cycle. Addressing the second point is more difficult–there may have been learning between LGAs that we cannot capture in our data. However, this interpretation would cast doubt on the incentive design of the project as the driver of improvements in state capacity, since the non-project LGAs did not receive these incentives but nevertheless improved over time.

These results raise further questions. What is driving the changes observed in state capacity in local governments that did not receive the project? What can we learn from this about how state capacity evolves over time? What is the role of external partners in the process? What can external partners do differently than the project we have evaluated to bring additional value to within-country processes of endogenous change? The rich survey data we gathered allows us to offer some answers to these questions, along with ideas for how external partners might innovate for greater impact.

Prior research steers us towards looking at the role of citizens for an explanation of how improvements came about in non-project local government areas. Available research has argued

that state capacity is built over time as societies become more complex and demand public goods that only a state can provide, such as defense against external aggressors (Tilly, 1992; Besley & Persson, 2009). The driving force of state capacity in this research is citizen willingness to pay taxes to a state they regard as legitimate (Besley, 2020; Fergusson et al., 2020; Weigel, 2020). Consistent with this research, the survey data show improvements over time in citizens' reports of actual payment as well as willingness to pay taxes so that local governments can develop their capacity. That is, the survey data suggests that state legitimacy to collect taxes has improved over time, as reflected in greater reports of compliance with taxes. The pattern of citizen responses also exhibit improvements in trust that local governments will deliver services and listen to citizen demands. Importantly, there is no difference in these survey measures of increasing trust in and legitimacy of the local state across the project and non-project areas. The data thus suggest that improvements in state capacity in Tanzania stem from endogenous changes in trust and legitimacy in the country rather than as a result of financial incentives offered by external partners.

This interpretation can be illustrated by examining a particular project component which targeted processes of oversight by and accountability to citizens. One-fifth of the performance score of LGAs under the project is allocated to indicators of consultation with citizens and disclosure of budget information for accountability (World Bank, 2012). In fact, this accountability and oversight component is the one where almost all project LGAs, 17 out of the 18, achieved the highest possible score by the time of their performance assessment in 2016-17 (World Bank, 2018b). Yet, we find no difference between project and non-project LGAs in the 2016 survey (as we might expect if project communities had already improved dramatically and non-project communities had not), nor in changes between 2016 and 2018, in household responses to questions about knowledge of public budgets and consultation by the local governments. That is, when measuring citizens' self-reported knowledge of and engagement with local governments' initiatives, we do not find evidence that the project improved citizen oversight and accountability. Citizens in non-project areas reported similar increases over time in knowledge about local government activities, and project areas showed no increases in self-reported indicators of citizen participation (e.g., whether they contacted any government officials or participated in meetings).

The evidence of a lack of value-added of an external project to endogenous processes of change within the country has linkages with several different strands of the literature on state capacity. First, it adds to the growing literature on state capacity in economic development (Acemoglu, 2005; Besley & Persson, 2009; Besley & Ghatak, 2005; Acemoglu et al., 2015; Dal Bó et al., 2013; Muralidharan et al., 2016; Fergusson et al., 2020; Besley, 2020; Bisin, 2020; Bowles, 2020; Papaioannou, 2020). Research examining how the current high income countries of the world built state capacity has concluded that the impetus came from growing demand in society for public goods such as defense against external aggressors, and for municipal infrastructure during the Industrial Revolution (Tilly, 1992; Besley & Persson, 2009; Lizzeri & Persico, 2004). In the case of low income, developing countries, research has found variation within and across countries of historical institutions of state capacity which continue to have persistent impact on

contemporary outcomes, long after those formal institutions have disappeared (Bandyopadhyay & Green, 2016; Gennaioli & Rainer, 2007; Michalopoulos & Papaioannou, 2013; Dell & Olken, 2020; Dell et al., 2018; Dell, 2010; Lowes et al., 2017).[5] These persistent effects of history suggest that state capacity takes time to build because it involves not just physical or concrete investments in recruiting personnel, collecting taxes and enforcing compliance but because it needs norms to evolve within societies (World Bank, 2016c, 2017; Khemani, 2019).

Second, the paper provides an empirical test relevant to prior qualitative critiques of development aid (Bourguignon & Gunning, 2018; Andrews et al., 2013, 2017; World Bank, 2017). Critics argue that development agencies focus on building formal institutional capacity in the image of developed countries' state institutions, which may result in developing countries "looking like a state" but lacking real state capabilities (Pritchett et al., 2013). The evidence we find can be regarded as consistent with this critique, although the project did try to go beyond transplanting formal institutions, and into areas of citizen oversight and accountability. In general, the paper contributes micro-empirical evidence to a large cross-country literature on aid effectiveness (Rajan & Subramanian, 2007, 2008; Bourguignon & Sundberg, 2007; Bourguignon & Platteau, 2015; Brautigam & Knack, 2004; Knack, 2001).

Third, the paper links to a growing body of evaluations (many of them randomized controlled trials, or RCTs) of policy interventions on how to build state capacity. For example, one study provides evidence on how to successfully recruit state personnel (Dal Bó et al., 2013) and another shows how technology can be used to better manage state finances and establish a "leakage-free" payment infrastructure (Muralidharan et al., 2016). Indeed, because of growing evaluations in this field, there is more evidence available to policymakers about what concrete policy actions to pursue than ever before (Banerjee & Duflo, 2012). The open questions pertain to the incentives of policymakers who have the power to take up such evidence and make policy choices and investments in state capacity on its basis (Hjort et al., 2019). This paper examines whether external development agencies can create these incentives through their projects, grants and loans. Our results of no difference in changes over time between project and non-project local government areas suggest that, at least in this context, the financing incentives provided by external partners did not add significant value beyond processes of change already enfolding in the country.

Fourth, the paper links to research on decentralization or the role of devolving powers to locally elected governments (Bardhan & Mookherjee, 2000, 2006, 2016; Faguet, 2003, 2014; Devarajan et al., 2009; Khemani, 2015). State building across many developing countries, especially those afflicted by conflict, has focused on locally elected leaders who may have information about and standing in their communities to develop trust and legitimacy (World Bank, 2011; Myerson, 2011). The project we evaluate in Tanzania is part of a large portfolio of lending and grant-making by international development partners to strengthen locally elected governments (Independent Evaluation Group, 2018). The results we find suggest that national governments are able to invest in building the local state even in the absence of international aid incentives and conditionalities,

[5] Tanzania was reclassified as a lower middle income country in 2020, but at the time of this work, it was classified as a low income country (Battaile, 2020).

as research has found in other countries (Acemoglu et al., 2015). The results in the institutional context of Tanzania further link to research on how states governed by a single national political party or the military, with concentration of power at the center, choose to invest in locally elected governments to build their capacity to deliver services at the frontlines (Martinez-Bravo, Mukherjee, & Stegmann, 2017; Ferraz, Finan, & Martinez-Bravo, 2020).

Finally, the paper links to a large literature on institutions and development. Reviews of this literature all point to political institutions as key to bringing about state capacity by shaping the incentives of powerful policy makers to scale up concrete actions, such as those that have been shown to be effective in previous evaluations (World Bank, 2017, 2016c; Dal Bo & Finan, 2016; Khemani, 2019; Olken & Pande, 2013). The comparative advantage or value-added of external partners has been explored in this research and yields ideas for innovation and experimentation in development projects (World Bank, 2016c; Devarajan et al., 2009). More evaluation of projects in different institutional contexts is needed to understand the role of external partners in building state capacity. Previous research has also shown the importance of considering the legitimacy of locally-originated program versus those led by external partners (Dal Bó et al., 2010). More experimentation or innovation is needed, drawing insights from recent advances in research on norms, trust and legitimacy as the driving forces of state capacity (Khemani, 2020; Besley, 2020).

The paper is organized as follows. Section 2 discusses the concept of state capacity, how it is measured in the research literature, and how it links to the design of capacity building projects pursued by external development partners. Section 3 describes the data that was gathered to evaluate the impact of such a project, using the opportunity available in Tanzania. Section 4 presents the methodology and results. Section 5 considers various explanations for these results, forwards our interpretation and discusses associated caveats. Section 6 concludes by offering some recommendations for innovation in capacity building projects.

2 State Capacity: Theory and Measurement

Economics has long assumed the existence of basic state capacity to protect property rights, enforce contracts, and maintain law and order, as the fundamental institutional conditions which are needed for market-led economic development. This recognition of state capacity is above the debate about the size of government or where it should intervene.[6] The role of state capacity in economic development has come to the fore in recent research motivated by the observation that higher income countries have systematically higher tax to GDP ratios than lower income countries (Besley & Persson, 2009, 2011; Acemoglu, 2005). As Besley & Persson (2009) observe: "A striking feature of economic development is an apparent symbiotic evolution of strong states

[6] Fukuyama (2004) reviews the history of ideas over the decades of the 1980s and 1990s about the appropriate role of state intervention in markets. He draws a distinction between the debate over the scope of state activity from the consensus over the need for a strong state that is capable of implementing at least those minimum activities needed from a state even under limited government (such as the rule of law). Acemoglu (2005) provides an economic framework for examining the tension between the need for "limited government" and yet "strong states".

and strong market economies."[7]

State capacity has been measured in economic research primarily as the ratio of government tax revenues to gross domestic product (Besley & Persson, 2011). The tax to GDP ratio serves as a summary statistic of sorts of the ability of governments to raise revenues to invest in the protection of property rights and the establishment of law and order, what sociologists and political scientists have termed a "monopoly over violence" (Weber, 1946; Anter, 2020). The reach of the state into local areas is also measured by the ability of local governments to collect revenues and administer state policies (Acemoglu et al., 2015). The process of building state capacity involves investments by national governments in the ability of local government agencies to administer policies (Dal Bó et al., 2013; Muralidharan et al., 2016).

The driving force of the process of building state capacity has been identified as citizens' demand for public goods that only a state can provide; state capacity in developed countries has been explained as a result of citizens being willing to pay the taxes needed to finance public goods and the state institutions that would provide them (Tilly, 1992; Besley & Persson, 2009; Lizzeri & Persico, 2004). In the contemporary world where some countries are clustered around high income and high state capacity and others at the opposite end of low income and low state capacity, international development partners have assumed a role in building capacity in developing countries (Jones et al., 2006; Levy & Kpundeh, 2004). Furthermore, over the past three decades, the practice of international development has moved from financial transfers and policy advocacy as the primary way of effecting development to increasingly focusing on building institutions and country ownership (de Janvry & Dethier, 2012). Both practice and research have revealed that when state institutions are weak, which is too often the case, developing countries are unable to put external aid to effective use in growing their economies (Rajan & Subramanian, 2007, 2008). Since the 1990s, the World Bank, the largest development bank in the world, has designed projects to build state capacity both at national and sub-national levels of local government (Independent Evaluation Group, 2008). However, there is little research available on the impact of these capacity building projects.

The Tanzania Urban Local Government Strengthening Program (ULGSP) we examine contains essential features of how the World Bank has approached local state capacity building in its projects. For example, a project with the same features was undertaken in India and is described in World Bank (2016a). At the core of these programs are fiscal incentives based on assessments of institutional performance, termed the Annual Performance Assessments (APAs), for which the project mandates guidelines, scoring methodology, and the engagement of an independent third party (typically an accounting and audit firm) for its execution. The project also includes facilities for local governments to access training or advice on how to improve the outcomes measured by the APAs.

The rationale behind the design is that these performance grants will strengthen the incentives of local governments to undertake activities, and access the training needed, to increase

[7] Acemoglu & Robinson (2020) provide an extensive discussion of the interaction between state capacity and private actors along the development path of nations.

their scores, which in turn will be equivalent to building state capacity, as defined by the indicators in the APAs.

The specific indicators in the APA in the Tanzania project are:

(i) **Urban planning system:** documentation and indicators of having a General Planning Scheme in place.

(ii) **Fiduciary or financial management system**: documentation of internal audit reports undertaken by a fully constituted Internal Audit Committee, and scores on a system of public procurement.

(iii) **Infrastructure management**: documentation and verification of the utilization of financing to deliver physical infrastructure such as roads and sanitation services.

(iv) **Accountability and oversight by citizens**: verifying public disclosure of information about local budgets and convening of public meetings.

In addition to these four, a fifth area of own source revenue generation (from the property taxes assigned to local governments) was targeted in the original evaluation design:

(v) **Local property tax system**: increase in own source revenue from the collection of local property taxes. The goal was to enable local governments to gradually move away from dependence on fiscal grants to becoming self-reliant on own source revenue, of which local property taxes are typically the most important. However, in July 2016, three years into the program, the national government recentralized property tax collection, shifting it from local governments to the Tanzanian Revenue Authority (TRA). In response, after 2016 the program decided not to assign any scores to performance in generating own source revenues.

This evaluation draws on data from both the APAs and our independent surveys administered to government officials and households. Each of these sources has value. Changes in the APAs document whether project LGAs improved on the specific indicators that the project targeted. The survey data play three key additional roles. First, many of the survey indicators provide an independent check on data gathered via the APAs.[8] Second, the surveys with government officials complement the APA data with broader measures of governance quality. Third, the surveys with households in the LGAs provide a further measure of governance improvements, which is whether the public observes improved local governance.

A fundamental question for the impact evaluation is whether local governments that did not receive the project, and thus access to the incentives and training under the project, made similar improvements in the areas measured under the APAs. One obvious way to measure these outcomes in non-project local governments would be to engage the same firms that are scoring

[8]The survey data are independent in the sense that there are no rewards associated with the survey data, unlike the APA data which constitute the basis to allocate project funds to LGAs.

the project local governments on the APA to undertake the same process of data gathering and scoring for non-project LGAs. However, due to the intensity of data gathering required for the APAs under the project, the same auditing firm could not administer APAs in comparison LGAs under the same timeline as project LGAs. Given both this and the additional cost, the evaluation draws on the survey data for the difference-in-differences analysis. We also report single differences results using the APAs in project LGAs (see Table B1).

Through extensive field testing and in collaboration with the government to ensure we measured indicators of state capacity that were locally valued, we designed the surveys to gather the following types of data grouped into modules under the different categories covered by the APAs. In each module, we aimed to include at least some questions in accordance with the project guidelines on how to undertake the APAs, so that the measures are as close as possible to the outcomes incentivized by the project. In the list below, we discuss the consistency between the APA questions/scoring guidelines and the questions in our survey.

1. On urban planning systems:

Government officials were asked the following questions that are likely to be informative about institutional capacity for urban planning.

- Whether a General Plan had been approved since 2015, and conditional upon an affirmative answer, whether the official in charge of planning could show interviewers the plan. This set of questions follows the guidelines issued by the project to the firms that were contracted to undertake the APAs. Specifically, the APA scoring guide indicates to assign points if LGAs are compliant with steps including plan preparation process, data analysis and plan adoption and approval
- The respondents' estimate of the percentage of unplanned settlements in the LGA
- The respondents' view of whether the LGA experiences delays in receiving guidelines for preparing their budget. This question and the following one directly measure one of the key outcomes targeted by the project—to reduce delays in communication between the President's Office for Regional and Local Governments (PO-RALG) and the local governments.
- The respondents' view of whether disbursements from central government are timely
- The respondents' view of whether the budget was executed in accordance with expected results. APAs review planning and utilization of annual plans for development budget.

To illustrate how such data can be useful in evaluating the impact of the project on institutional capacity, we can examine whether project local governments are more likely to have a General Plan and are able to produce it when compared to local governments not targeted by the program; whether government officials in project local governments estimate a smaller share

of settlements are unplanned; and are more likely to report no delays in communication and disbursements from the central governments, and in budgets being executed in accordance with expected results.

Citizens were asked the following questions that are likely to be informative about their assessment of the quality of urban planning.

- Extent to which they think the local council guarantees good use of revenues (standard 5-point scale used for such survey questions)
- Extent to which they think the local council makes good investment plans
- Whether they have observed problems with local government

2. Fiduciary or financial management system:

Government officials were asked the following questions to assess both whether de jure internal audit systems are in place, and their views of how effective internal audits are in monitoring the use of funds.

- Whether all of the positions on the internal audit committee are filled. This is consistent with the measure in the APAs on whether audit committees are in place and operational.
- Whether the internal auditor can show the interviewer copies of internal audits, and how many of these. APAs review audit reports from previous fiscal years.
- Views on whether internal audits are independent of political interference
- Views on whether internal audits are effective in monitoring the use of funds
- Whether there has been turnover in the membership of the internal audit committee between the two survey rounds of 2016 and 2018. Again, APAs measure whether audit committees are in place and operational.
- whether there is a tender board in place, what its composition is and frequency of its meetings. APAs review existence and functioning of tender boards

Citizens were asked the following questions to measure their assessment of local corruption:

- The extent to which they trust the local council
- The extent to which they think councilors are honest in handling public money
- Whether they think most, some, or none of council members are corrupt
- Their experience with bribe payments for various services

3. Infrastructure management:

Government officials were asked the following questions relevant to their experience of managing infrastructure project implementation:

- Whether they think the number of engineers is adequate for the LGA's needs
- Whether they think payments to suppliers were carried out on time
- Whether measures have been taken to publicly disseminate information about the physical progress on infrastructure investments. APAs review whether development plans' progress is disseminated to the general public.

In the absence of data on the quality of infrastructure investments, drawn directly from measuring that quality at source (such as by taking samples of roads as in one project in Indonesia (Olken, 2007), or engineering assessments of sanitation infrastructure), the evaluation relies on citizen reports of the performance of governments in delivering urban infrastructure services:

- Citizens' assessments of whether the local government maintains roads well
- Whether the government keeps the community clean
- Whether the government manages land well
- Whether the governments maintains health standards well
- Ease of access to a variety of services, such as building permits
- Assessments of whether the neighborhoods in their ward are connected by paved roads, have garbage removed regularly, etc.

4. Accountability and oversight by citizens:

The following questions were asked of government officials:

- Whether there exists a formal mechanism for citizen feedback. APAs review whether procedures for dissemination and public participation are in place for the preparation and implementation of annual development plans, environmental and social impact assessments, and resettlement action plans.
- Whether there exists an official system to handle grievances
- Whether the respondent has handled grievances personally
- In how many public meetings have infrastructure investments been discussed. Again, APAs review whether development plans' progress is disseminated to the general public

The following questions were asked of citizens:

- Whether the local council provides information on budgets
- Whether it is easy to find out their tax bill
- Whether it is easy to find out how revenues are used
- Whether it is easy to find out how the LGA spends revenues
- If the local council allows participation
- If the local council consults before decisions
- How the local council handles complaints
- If an ordinary person can affect local governance
- Contact with government officers
- Participation in meetings
- Awareness of grievance process
- Whether local leaders (mayor, councilors, community or mtaa leaders) try to listen to citizens[9]

5. Local property tax system:
Government officials were asked the following questions:

- Whether some form of incentives to citizens to increase tax revenues had been tried.
- Whether the respondent pays property tax on residential/commercial buildings. This question may be a particularly good measure of the ability of the state to collect local taxes, regardless of whether the collection is administered by the national revenue authority or the local government. A government report on the challenge of domestic revenue mobilization from property taxes has identified the recalcitrance of local politicians, who tend to be property owners, as a problem (Government of the United Republic of Tanzania, 2013). Hence, whether local officials have paid their property taxes is an indicator of the fiscal power of the state, the key measure of state capacity in the research literature.
- What percentage of tax invoice is collected.
- The percentage of total properties registered.
- Whether the respondent agrees there are opportunities to raise local revenues
- Whether the respondent agrees that the challenge to raising local revenue is political

[9]The mtaa is the smallest administrative unit in Tanzania. In rural areas, a mtaa may be a village. In urban areas, a mtaa may be a neighborhood.

The project, and so the APAs, dropped the component on local property tax systems in 2016 because the national government re-centralized the collection of property taxes to the Tanzania Revenue Authority. Citizens were asked the following question:

- Willingness to pay taxes, as measured by the extent to which they agree with statements along the following lines: citizens should pay taxes so the local government develops; better to pay high taxes to get more services; not paying taxes is wrong and punishable; tax code and collection is fair. These types of questions are the focus of current research on state capacity (Besley, 2020; Papaioannou, 2020; Bisin, 2020; Bowles, 2020).

An agnostic approach to measuring the impact of the project involves examining all the variables described above as equally likely to be important for state capacity and allowing the data to reveal where there are any differences in changes over time across project and non-project local governments. As we will discuss in detail below, a striking finding from the data is that only one of the many variables listed above exhibits greater improvement over time in project than in non-project local governments. We will also focus on the outcomes emphasized in the current research literature as measures of state capacity—the ability to raise revenues. Although the project dropped the component on local property tax systems in 2016 because the national government re-centralized the collection of property taxes to the Tanzania Revenue Authority, if state capacity is increasing in the country over time, we would expect to see this captured in the questions listed above on willingness to pay taxes to contribute to the development of the local state.

In addition to the variables discussed so far, our surveys included additional modules to gather data on the administrative capacity of local government officials such as whether they have access to computers, can write emails and memos, and have received any training recently. We also estimate the impact of the project on these measures of administrative capacity of local personnel, in line with how the literature has examined local state capacity (Acemoglu et al., 2015).

Finally, other modules in the survey drew upon advances in research on the importance of informal institutions, such as culture or norms and beliefs in complex organizations, which shapes their capacity to perform (Bloom & Van Reenen, 2010; Rasul & Rogger, 2018). Khemani (2019) provides a review focusing on public sector organizations. These modules included the following types of questions posed to local government officials:

- Extent to which officials feel peer pressure to perform their tasks well
- Extent to which officials take pride in their work
- Extent to which they trust their peers
- Extent to which they share values with their peers

In sum, the outcome variables on which impact is evaluated are numerous and rich, with careful surveys of two types of respondents—local government officials and households. These

outcomes include all the areas targeted by the project in its assessment of institutional performance—urban planning, fiduciary systems, infrastructure management, revenue generation and accountability and oversight by citizens. In addition, we examine impact on measures of state capacity emphasized in the growing research literature—ability to raise revenues; administrative capacity of state personnel; and culture and norms of performance in the organizations of local government.

3 Data

3.1 Data sources and timeline of data collection

Two rounds of data were collected as part of the impact evaluation. We present a timeline of the surveys' implementation and the APAs in Figure 1. The first survey was completed in February 2016 when some project interventions had been implemented in most LGAs. As such, the first survey can be considered as capturing the situation at early stages of project implementation. Likewise, the second survey was conducted in April-May 2018, when not all project activities had been completed, and therefore can be considered as capturing late stages of project activity. Specifically, the temporal distance between the first and second survey rounds is of 26 months which is around 45% of the temporal distance between the 1st APA and the 6th APA (≈60 months). During this period, about half of the project funding commitment was disbursed.[10] As part of the project, APAs were also collected at regular intervals.

3.2 Panel data

The data used for this study are built into two panel datasets (i.e., two datasets with data from the same LGAs and/or households at two different times): one of households and one of government officials. Households and government officials were interviewed in all 40 original evaluation LGAs (both in the first and second rounds) and in 12 newly added LGAs (only in the second round). The 40 LGAs consisted of 18 project LGAs and 22 comparator LGAs. The process of selecting comparator LGAs, drawing on a mix of propensity score matching and local expert surveys, is detailed in Appendix A. Since the main goal of this paper is to assess changes in outcomes between the first and second rounds of surveys, for the remainder of the text we will only refer to the 40 LGAs originally present in the first round survey, ignoring the additional LGAs added at the second round.

The Household panel dataset includes 5,996 observations across the two waves - 2,998 households were interviewed in each wave. The sampling of households in the first survey round was

[10]This timing could present a problem if all the gains in the program took place at the very beginning, before most activities were implemented, and/or at the very end. According to the APAs, indicators kept growing over subsequent APAs which suggests that not all gains were achieved at the early stages and indicators do not present large jumps towards the end of the period, i.e. between 5th APA and 6th APA.

performed in the following manner: in each of the 40 LGAs, 3 wards were selected[11] and then one enumeration area (based on the census) was sampled in each ward. All households living in that enumeration area were listed and 25 were randomly selected to be interviewed, for a total of 3,000 households. Around 80% of those were tracked in the follow-up survey and re-interviewed. For those not found, a new household was interviewed as a replacement.[12]

The Government Officials panel dataset includes 948 observations - 474 individuals in each of the waves. In the first round, the following 12 government officials were targeted to be interviewed in each LGA: Mayor/Council Chairperson; Council Director; Council Internal Auditor; Council Economist; Council Human Resources officer; 3 Elected Ward Councilors; and 3 Ward Executive Officers. Government officials were assured their answers would be kept confidential, and in 95% of cases the respondent was alone for the entire interview. At the second survey, over 70% of officials still worked at the same LGAs and were re-interviewed even if working in a new position. For the remaining officials, the interview was conducted with the individual currently occupying the position of the respondent in the initial survey.[13] For most of these officials, PO-RALG oversees all decisions related to appointments, transfer, promotions, etc. As PO-RALG also implemented the project and commissioned the survey, this could create some bias in government officials' responses. However, this potential bias would affect both treatment and comparison groups.

For both households' and government officials' surveys, our main difference-in-differences specification compares average changes in outcomes in project vs. non- project respondents, pooling together respondents that participated in both waves and replacements for those who could not be tracked. Our main estimates, in other words, treat the panel dataset as two separate cross-sectional surveys.[14]

3.3 Outcomes of interest and aggregate indices

As discussed above, the survey includes indicators of institutional capacity among officials and citizens in several dimensions that can be directly related to the five core systems targeted by the project: the urban planning system, the fiduciary or financial management system; infrastructure management; accountability and oversight by citizens; and the local property tax system. Accordingly, we structured the outcomes of interest, in both the household and government official surveys, in those five areas.

[11] Wards were selected in a manner that permits a balanced distribution between areas that received infrastructure projects and area that did not. Specifically, the ward in which the LGA headquarters is based was always selected and then one ward with a recent infrastructure project and one without a recent infrastructure project were randomly sampled. No sampling weights were included in the analysis as our ward sampling strategy should produce a sample representative of the LGA level.

[12] Attrition of original households was higher in ULGSP LGAs: 20% of households were replaced in those regions vs. 15% in control LGAs.

[13] The share of officials still holding the same position at endline was not statistically different between ULGSP and control areas (67% in ULGSP vs. 62% in control LGAs)

[14] Restricting the sample to only those households/officials that were interviewed in both surveys and estimating a fixed-effects panel model does not significantly change the results.

In what follows, we will present results using questions that aim to evaluate responses of state capacity in each of those areas. In addition to presenting results for an exhaustive set of individual outcomes, we also construct indices that aggregate those outcomes for each system. The construction of these indices follows closely the methodology in Anderson (2008). For each index, we first code all components so that higher values indicate a "better" outcome, then standardize all transformed variables and finally construct the index as a weighted average of components.[15]

The indices for the household survey and their underlying components are as follows:

- **Government officials' survey:**

    1. *Staff Capacity Index* (6 items): Responses on staff skills in using computers and software.
    2. *Management Index* (5 items): Responses on management willingness and ability to attract, retain, and promote staff.
    3. *Performance Culture Index* (8 items): Responses about staff shared values, commitment to deliver, pride in serving, and ability to withstand political pressure.

- **Household survey:**

    1. *Urban Planning Systems Index* (3 items): Responses about local government use of revenues and planning.
    2. *Fiduciary Responsibility Index* (8 items): Responses about government honesty and bribe payments.
    3. *Infrastructure Management Index* (12 items): Views on government capacity to maintain infrastructure and access to state services.
    4. *Accountability & Transparency Index* (21 items): Views on local government transparency, participation in political meetings, ease of access to information.
    5. *Views on Taxation and Fees Index* (11 items): Normative views on taxation and fees (e.g., is it wrong not to pay taxes?) and positive views about whether individuals are punished for not paying fees and taxes.

In Table 1 we present correlations across the summary indexes from the household survey on government performance in different areas. All the measures are broadly positively correlated: positive responses or evaluations in one dimension are usually accompanied by positive evaluations in similar dimensions. It is also clear, however, that these measures are not perfectly correlated, i.e., they likely capture different dimensions of citizens' interaction with the state.

[15] Weights for each component are equal to the sum of row-entries in the inverse covariance matrix. That means the index underweights components highly correlated with other components and over-weights those with "new" information–i.e., those with a low correlation with other components.

The strongest correlation is between the Infrastructure Management index, which reflects quality of services, and the Accountability and Transparency index, which focuses on ease of access to information and transparency, at 0.33.

We also include data on the relationship between the survey measures and the APAs, in Appendix B. Since APAs were only collected at ULGSP LGAs, the sample is restricted to 18 observations. While for the household survey we observe overall positive correlations between outcomes measured at baseline, endline and changes over time using survey and APA indicators (Figure B3 and Figure B4), results for the government official survey are more mixed, with overall smaller correlation in magnitude and both positive and negative relationships (Figure B5 and Figure B6). We should treat these correlations with caution, not only due to the limited number of observations but also because in some indicators the amount of variation is quite limited. For example, almost all ULGSP LGAs had scored the maximum amount in the Oversight component of APA by our first-round survey, so variation between our first and second surveys is zero for the majority of them).

3.4 Evidence of successful project implementation

An important initial question regards the implementation of the program: does the survey provide evidence that government officials were aware of the program and that they received training and funding as expected? The answer is a definitive yes: in the first round, officials in project LGAs recognize the project as the main source of capacity training and budget support for infrastructure. In Figure 2, panel A, we show that among officials in project areas, 65% report the project as the agency most supporting capacity building vs. 12% in comparison LGAs. For comparison LGAs, the main agencies reported as supporting capacity building are sector ministries (30%) and community-based organizations (18%).

Regarding budget support for infrastructure, panel B of Figure 2 shows that 50 percent of officials in treated LGAs choose the project as the main source for recent increases in budget for infrastructure vs. 0 percent in comparison. Twenty-six percent of project officials and 46 percent of non-project officials indicate own revenues as the most important source of finance, whereas 41 percent of officials in non-project LGAs indicate government grants as the most relevant source.

As discussed above, the first survey round was fielded after the 3rd APA, when government officials would have been aware of the project. The data above suggest that officials were not only aware of the existence of the program, but also report the project as being the most relevant source of capacity building and new finance for infrastructure. At the same time, panel B of Figure 2 shows that project LGAs still substantially rely on government transfers and own source revenues to finance infrastructure, which suggests that a redirection of central government resources from project to not-project LGAs is not happening, or at least not on a large scale.

3.5 Evidence of systematic observable differences between project and comparison LGAs

As previously discussed, while the project was not randomly assigned to LGAs, the government and the evaluation team selected comparison LGAs that were likely to be similar to the ones receiving the program across a range of observable indicators.[16] This matching happened before the fielding of the surveys, however it is not obvious that respondents in project and comparison LGAs would be identical in the main indicators given the targeting of the program. The key assumption behind our estimation stategy is not that project and comparison LGAs be identical but rather that they be developing at similar rates in the absence of the project (often called the "parallel trends" assumption). In this section we characterize the project and non-project LGAs and provide evidence supporting the parallel trends assumption.

We perform this comparison for an exhaustive range of indicators in Table 2 and Table 3, for households and officials, respectively. The results suggest that respondents in project areas are, as expected, consistently different from those in non-project areas. In both tables, the first and second columns present the average value of the indicator for project and non-project areas, respectively. The third column presents the difference between those means, while the fourth presents the p-value of a T-test of mean equality. The last column reports the number of respondents with non-missing values for each test.

The first panel of Table 2 documents that respondents are demographically different in ULGSP and comparison areas: those in project areas come from smaller households; are 10 p.p. more likely to be literate and 15 p.p. more likely to have more than complete upper secondary education; and are 30 p.p. less likely to work in agriculture. Overall, these demographic characteristics suggest that households in project areas are wealthier than those interviewed in non-project LGAs. This is confirmed by answers on asset ownership: respondents in project areas are 4 times as likely to own a car as those in comparison areas; almost 3 times as likely to own a TV, 6 times as likely to own a computer and more than 2 times as likely to own a refrigerator.

Not only are demographic characteristics very different between project and comparison areas, but responses on government performance and accessibility to services are also consistently better in project areas. While 62 percent of respondents in those areas agree that the government maintains the roads well, only 47 percent of respondents in comparison areas agree. The gaps are smaller but still meaningful and statistically significant for responses about health standards, cleanliness and land management. Respondents in project areas are also consistently more positive about ease of accessibility to services: they are 8 percentage points and 12 percentage points more likely to agree that it is easy to access building permits and household services such as water, respectively.

Differences among government officials are less stark, but they still consistently suggest that project LGAs are better performers than comparison LGAs (Table 3). When asked whether the master plan is updated, for example, almost half of government officials in project areas answer

[16] All the details of the selection of the comparison group are discussed in Appendix A.

positively vs. 27 percent in comparison areas. Possibly directly related to the efforts of the project, almost all officials in those areas say the LGA has a formal plan for capacity building vs. 62 percent in comparison areas; and conditional on having a plan, less than half of respondents in comparison areas could show the plan vs. three-quarters of respondents in project LGAs. While these differences are statistically significant, we do not observe significant differences in other indicators such as reporting of budget preparation delays, budget execution or share of unplanned settlements. On the topic of fiduciary systems, project areas also perform better: respondents are 16 percentage points more likely to report having the internal auditor position filled, and they are more likely to agree that the internal audit office is independent and effective in monitoring. We do not observe significant differences, however, on most indicators under the topics of infrastructure management and accountability and transparency.

Taken together, these results suggest that between comparison LGAs and those receiving the project, systematic differences exist in the initial survey between respondents. Given the project's targeting rules, this is an expected result. Our identification strategy, a difference-in-differences approach, is well placed to estimate the causal effect of the program under pre-existing differences between treatment and comparison groups. We assess the main identification assumption behind a difference-in-differences approach by checking for parallel trends in state capacity levels before the start of the program. Specifically, we examine trends before the first round of data collection on a key measure of local state capacity, LGA's own source revenues. An important limitation here is that we only have available own source revenue data at the LGA level for the years 2007 to 2011 and data is available for the entire period of 5 years for only 36 out of the 40 LGAs (16 project and 20 comparison).

In Figure C1 we plot the mean per-capita revenues for both project and comparison LGAs. (We normalize for LGA population size in 2011.) The trends follow a very similar path between 2007 and 2009, starting to show some divergence after 2009. Given the paucity of data, we cannot run a more thorough exercise which would include an event study (including the periods before and after the start of the program). However, when we run a simple statistical test of the difference between the mean per-capita revenue in project and comparison groups, we cannot reject the null of equality of the means.[17]

To further explore the parallel trends assumption, we also use nighttime lights data. Nighttime lights data are highly granular and are available over a long period of time and so enables us to complement the pre-trend analysis of revenues. We focus on the period to 2000 to 2018.[18] However, while a good proxy of local economic activity,[19] nighttime light data might not be a good proxy for local state capacity. With this important caveat in mind, we show that trends in nighttime light are remarkably similar over the period 2000 to 2013 (which is when the program came into effect) between project and comparison areas (see panel (a) in Figure C1). This find-

[17] Coefficients obtained estimating equation $OwnPercapitaRevenue = \alpha + \beta(ULGSP) + year + \epsilon$ with standard errors clustered at the LGA level.

[18] Details are in Appendix C.

[19] See Gibson et al. (2021) for a recent review.

ing is confirmed by the event study we run and show in panel (b) of the same figure. A similar conclusion can be reached if we look (again in Figure C1) at the time period from 2000 to 2016 (which is the year of the first round of data collection) or if we focus on the treatment period of 2016 to 2018.

To produce Figure C1 we focus on the 3 wards that were sampled in each LGA for this study.[20] In Appendix C we include graphs on different samples which summarize different ways of computing the average nighttime lights for each group (see Figure C2, Figure C3, and Figure C4). Overall, these further explorations suggest that trends are not drifting apart between the evaluation groups. At the same time, there is some indication that comparison areas might be catching up with treatment areas in terms of local economic activity over time.

4 Empirical Strategy

As discussed above, our goal is to assess whether LGAs that received the project improved more between the first and second survey rounds when compared to non-project LGAs. For both the government officials and household surveys, we use responses at the individual level and, in order to formally test for different improvement rates, estimate a difference-in-differences model of the form:

$$Y_{ilt} = \alpha + \delta \mathbb{1}\{ULGSP\}_{ilt} + \gamma \mathbb{1}\{SecondRound\}_{ilt} + \beta \mathbb{1}\{ULGSP\}_{ilt} * \mathbb{1}\{SecondRound\}_{il} + \Pi X_{ilt} + \epsilon_{ilt} \tag{1}$$

where Y_{ilt} is some outcome of interest of individual i in LGA l and period t; $\mathbb{1}\{SecondRound\}_{ilt}$ is an indicator that takes the value 1 if the respondent belongs to the second survey round in 2018, and 0 otherwise; $\mathbb{1}\{ULGSP\}_{ilt}$ is an indicator that takes value 1 if the respondent resides in a project LGA, and 0 otherwise; X_{ilt} are individual/household characteristics used as comparisons (welfare index, household size, age, gender, marital status, and education levels) and ϵ_{ilt} is an error term.[21]

To evaluate the hypothesis that project LGAs were differentially affected by the program, we formally test whether $\beta = 0$; the coefficient on the interaction between the indicator for 2018 and project LGAs gives us the difference between *changes* in project vs. non-project LGAs between 2016 and 2018 (i.e., this is the difference-in-differences coefficient). In our results, we also often present the estimates for the coefficient γ, which indicates the change in outcomes between 2016 and 2018 for the non-project LGAs.

In estimating this difference-in-differences model, we can only interpret the resulting estimates as causal if, in the absence of the intervention, the trends in outcomes for comparator and project LGAs would have been the same. In other words, for any given outcome (trust in local council, for example), our assumption is that were the project not to be implemented, the average

[20] The averages shown in the graph for each group are then the average of the wards' nighttime lights averages.

[21] Since we estimate the model at the individual level, but the treatment is at the LGA level, our estimates of standard-errors are clustered at the LGA level.

change in that indicator for LGAs that actually received the project would have been the same as that observed in LGAs that did not receive the program. Some evidence on the parallel trend assumption was provided in the section above.

4.1 Results

We now present the main results of the paper: did LGAs receiving the project present a different trend in indicators of state capacity when compared to the comparator areas? We highlight the trends within project areas, then present how the comparator areas perform, and finally compare the performance between treatment and comparator areas to explore the impact of the program.

4.2 How did outcomes change within ULGSP LGAs over time?

We start by presenting how the responses of households about institutional capacity changed in project areas. Many indicators are presented in Table 4, Table 5 and Table 6. Columns (1) and (2) present average responses among households in project LGAs, in 2016 (first round) and 2018 (second round) surveys, respectively, while column (5) presents the difference in means.

Overall, the response improved across the board. In Table 4, we see the share of households reporting that the local council makes good use of revenues has increased from 44% to 66%, for example, while the share saying that local governments make good investment plans increased by 21 p.p. Responses about government performance in delivering services such as road maintenance and cleanliness have also improved, as have responses about ease of access to building permits, household services and medical treatment.

Table 5 reports indicators related to responses of transparency and accountability. The share of respondents agreeing that the local council provides information on budget, allows participation, consults other actors before decision. and handles complaints well have all increased by over 10 p.p. - for all these indicators less than 60 percent of respondents agreed with the statements by the second survey, but the performance increased significantly in the two years between surveys. It is also worth noting that direct political participation, by contacting officials or participating in meetings, are both low and do not seem to be increasing: less than one in five respondents ever contacted a village official, a number that remained unchanged between surveys, and only 40 percent ever participated in village meetings.

Finally, Table 6 presents results related to response on taxation and preferences over public good and government decisions. We do observe an increase in the share of citizens reporting that citizens should pay electricity and that not paying electricity is wrong and punishable, but only half of respondents say that not paying taxes is wrong and punishable, a number that does not change in 2018. We do see, however, a large increase in the number of individuals agreeing that the tax system is fair.

Beyond households in project areas, both objective measures and views reported by government officials also improved. In Table 7, the share of respondents reporting an up-to-date master

plan increased from 46% to 69%, for example. It is also remarkable that, despite starting at a high level, the share of project areas with a formal plan for capacity building attained 100 percent by the second survey - most likely a direct result of the intervention. The share of respondents agreeing that the internal audit office is independent and effective both increased significantly between the two rounds.

Overall responses about capacity building activities also improved, as did opinions about staff capacity and norms: the share reporting that employees trust one another, share a strong set of values and take pride in their duties increased by 34 percentage points, 19 percentage points, and 23 percentage points, respectively. This broad range of improvements over time is consistent with the fact that APA measures also improved over time for project LGAs (Table B1).

4.3 How did outcomes change within comparison LGAs over time?

The previous section documented significant improvements in the responses of households and officials about state capacity in project areas, as well as objective measures such as having master plans up to date and auditor positions filled. We cannot, however, jump to the conclusion that these changes were caused by the program: they might have happened even in the absence of the intervention, if state capacity was in a trajectory of improvement. To assess whether impacts can be attributed to the project, we now present what happened in LGAs that did not receive the program.

The results are presented in the same Table 4, Table 5 and Table 6. Columns (3) and (4) present the average for each indicator among non-project respondents, for the first (2016) and second (2018) rounds, respectively. The difference in means is presented in column (6).

Column (6) suggests that overall improvements in responses were also observed among households and officials in comparator LGAs. Despite remaining at lower levels than those in project areas, the share of respondents that agree that the local council makes good use of revenues has increased from 38% to 57%, for example, and those that believe the government makes good investment plans increased from 42% to 64%. Responses about the government service delivery have also improved, as have opinions on ease of access to services: the share of individuals reporting easy access to medical treatment, for example, increased by 15 p.p. Some indicators did deteriorate, notably the share of individuals reporting that councilors are honest in handling public money and reporting easy access to waste collection. But overall responses about government performance and ease of access improved across the board.

That was also true, as reported in Table 5, for a wide range of indicators on accountability of government (such as access to information on budget and handling of complaints). The share of respondents reporting that it is easy to find out what taxes they need to pay doubled to 16%, as did those responding that it is easy to find how the LGA spends revenues. Objective measures of political participation, like contacting officials or reporting the existence of village meetings, recorded small decreases.

Finally, as in project villages, Table 6 shows that normative responses about paying taxes

and electricity also improved between the two survey rounds: the share of individuals that agree that citizens should pay taxes so local government develops increased from 41% to 50%, and those who think it is wrong not to pay electricity improved by 9 p.p.

Similar improvements were observed in the survey of government officials, presented in Table 7 and Table 8. The share of respondents affirming the LGA has a plan that is up to date improved from 27% to 44%; the share reporting a plan approved in the last 2 years increases by 26 p.p.; and the share of LGAs with formal capacity building plans increased by 30 p.p. Improvements were also observed in responses about budget delays and execution, agreements with independence of internal audits, and responses on both staff capacity and staff morale.

4.4 The impact of the project on household responses and government official reports of institutional quality

In order to estimate the causal effect of the project on areas that received it, as discussed above, we use areas that did not receive the program as counterfactuals. Under the assumption that these areas are valid counterfactuals (i.e., they would have followed similar trajectories in the absence of the program), assessing whether the project had an effect on the outcomes of interest is equivalent to examining whether areas that received the program had a differential change in outcomes, when compared to the comparison areas. The simplest (and most transparent) way to make this difference-in-differences estimation is to compare columns (5) and (6) in Tables 4 through 8: were the changes in project areas larger than those in comparison areas? Simple visual inspection suggests that, overall, this is not the case: project LGAs do not seem to consistently outperform comparison areas, which often presented larger improvements.

To formally test whether changes in project areas were different from changes in comparison areas, we estimate equation 1 above, including a series of individual level controls.[22] Figure 3 - Figure 5 present the results for these regressions for the household survey, while Figure 6 and Figure 7 present results for government officials. The left panel in each figure presents the point estimate and 95% confidence interval for the coefficient γ, indicating the change in outcome,[23] between 2016 and 2018, for comparison LGAs. The right panel presents the same for coefficient β, and it is the difference-in-differences coefficient: it represents the differential change in outcome among project respondents, when compared to non-project respondents.

Focusing first on the left-hand panel, we observe that for a vast number of indicators, as discussed above, comparison areas recorded improvements between 2016 and 2018. All indicators are constructed such that higher values indicate normatively better outcomes or improved views, and very few estimates indicate worsening performance between first and second rounds, in

[22] Controls include household size, age, gender, marital status and education level. Note that we are not estimating individual fixed-effect models, so we can include time-invariant comparisons. Furthermore, as discussed above, some households and officials are replaced in the second round survey so controls would still vary over time for some units. All standard errors are clustered at the LGA level. Due to the small number of clusters, we also report 95% confidence-interval for our estimates using wild-bootstrapping. Results are not sensitive to the inference method.

[23] For ease of reading, all dependent variables were standardized before the regressions, so coefficients are to be interpreted as changes in standard-deviation units of the dependent variable.

either survey. From the household survey, trust in the local council did fall, as did the share of respondents reporting that most neighborhoods are accessible by road and that had satisfactory responses from officials regarding complaints. Among government officials, no results suggest worsening response about state capacity, and several indicators related to planning systems, taxation, capacity building and culture of performance registered improvements.

Turning to the right-hand panel, the observation is that for almost all indicators, changes in project areas were not different from those observed in comparison areas (i.e., estimates of the β coefficient are often very small in magnitude and statistically indistinguishable from zero). Among households' responses about state capacity, for example, no indicators related to urban planning, fiduciary systems, accountability or taxation changed differentially in project areas between the first and second survey rounds. The only indicator that recorded a differential change in the treated areas was the share of neighborhoods accessible by road, which decreased in in comparison areas while increasing in project LGAs.

Among government officials, the same pattern repeats: for the vast majority of indicators we observe no differential trend in project vs. comparison LGAs. For the few indicators that did change differentially, we observe better performance of comparison LGAs: for several indicators related to taxation, like seeing opportunity to increase tax revenue and respondents paying taxes on their own property, performance improved more in comparison than project areas.

We also present results for the aggregate indices created, in Table 9 and Table 10. In both tables, panel A presents results for a simple difference-in-differences specification, without including any comparisons, while Panel B includes individual-level comparisons. For both panels, Table 9 shows that, except for the Accountability and Oversight index, respondents in comparison LGAs were more positive in the second survey than in the first one - as documented by the positive coefficients of the indicator variable for the 2018 survey round. The improvement in project LGAs, however, was no different than that observed in comparison LGAs: the coefficients on the interaction between the project indicator and the second round survey indicator is small and indistinguishable from zero. Across all outcomes, we can reject impacts larger than 0.3-0.4 s.d. [24] This is consistent with the fact that we observed no differential improvement in responses in the vast majority of the underlying variables used to construct these indices.

The same overall result is found for the government officials' survey on Table 10. Here we present three aggregate indices - the staff capacity index, the management index, and the performance culture index - as well as three of the main indicators of state capacity - having an updated master plan, perceiving the internal audit office as independent, and capacity to raise local revenue. For all outcomes, the point estimates suggests an improvement in the second survey round, though not always statistically significant. The difference-in-differences coefficients, on the other hand, are always smaller in magnitude and never statistically different from zero. For several estimates, point estimates are negative, suggesting that, if anything, comparison LGAs

[24] As a reference, we estimate overall increases of 0.41 and 0.52 s.d. for the indices on Urban Planning system and Views on taxation and fees, respectively, between the first and second round surveys. In both indices the difference-in-differences estimates are an order of magnitude smaller, and we can reject effects larger than 0.25 s.d.

outperformed project areas. For the standardized indices, we can reject differential improvement in project areas as small as 0.3 standard deviations (Management), 0.4 s.d. (Staff capacity) and 0.5 s.d. (Performance culture). Once again, we find no evidence that project LGAs registered a larger improvement in responses on performance when compared to comparison LGAs.[25]

5 Discussion

In this section, we consider possible explanations for the lack of statistical difference in these outcome measures and what these results imply for the design of external capacity building projects going forward. First, we observe a pattern of positive changes over time in the responses from government officials about their experience with managing local responsibilities and delivering services, and from citizens about receiving these services and being willing to pay taxes to finance them, in both project and non-project LGAs. That pattern is consistent with overall change in Tanzania in the direction of strengthening state capacity.

Second, the additional (besides country-level processes of change) value of the project's specific activities to strengthen incentives of local government officials—such as through the Annual Performance Assessment (APA)—appears to be low, with no evidence of difference in responses of government officials across project and non-project areas. There is also no evidence that the project made a difference for the citizen oversight and accountability channel through which the project sought to strengthen incentives of local government officials. Citizens in non-project areas reported comparable increases over time in knowledge about local government activities, and project areas showed no increases in concrete indicators of citizen participation (such as whether they contacted any government officials or participated in meetings) despite the APA giving the highest possible score to project LGAs on this component.

What alternative explanations could fit the pattern of results that we observe? First, it is possible that the lack of difference in project and non-project areas could be due to project LGAs demonstrating superior performance and thus having other LGAs in the country learn from and copy their practices. It is also possible that investing in local state capacity is a strategic complement across local governments (Acemoglu et al., 2015), such that as state capacity increased in project areas, other LGAs perceived greater returns from investing in their own capacity. However, the complete lack of any statistical difference across a rich set of variables casts doubt on these explanations. If the project had such large demonstration effects, we would expect to see at least some difference in some of the variables, rather than have all the benefits spill over in this short span of time. Even if LGAs were learning from each other, the significant improvements over time in non-project LGAs, which did not receive the project's incentives or capacity building activities, cast doubt on the mechanism design of the project–of incentives generated by performance grants.

[25]We present alternative specifications to the standard difference-in-differences estimator – fixed effects and a semiparametric difference-in-differences estimator – using government officials and households in Appendix D: Table D1 and Table D2. Estimates are qualitatively similar and suggest an overall null impact of the intervention.

Second, the project may have had its principal impact before the first survey in February 2016, in which case the difference between the two surveys might not capture project-induced improvements. We examine the Annual Performance Assessments (APAs) of the project to discern which components show the greatest increase in project measures of institutional performance. In Figure B1 we present the mean scores across project LGAs in each of the performance assessments. Our first round survey was implemented in early 2016, around the fourth APA, while our second round survey happened in 2018, close to the sixth APA. We find that the "Accountability" component of the APA shows a large increase between the 2nd and 3rd APAs, while the other components (Revenues, Infrastructure and Urban Planning) do not show a pattern of concentrated growth before our first survey round. (We discuss the available evidence in more detail in Appendix B.) It could be that had we undertaken a citizen survey in 2013, we may have found that the 18 project LGAs had much lower levels of citizen-survey-based Accountability measures than the 22 non-project LGAs in 2013? What if the project incentivized the 18 LGAs who were reluctant to publicly post their budget information in the absence of these project incentives; what if our 2016 survey measure, compared to this hypothetical 2013 survey measure, shows that Accountability significantly improved, as a result of the project incentivizing the LGAs to reach out to citizens? We cannot rule this out because we do not have survey data from 2013 across both project and non-project LGAs. We find no significant difference between project and non-project LGAs in the 2016 survey measures of citizen engagement targeted by the Accountability component, nor in changes in citizen engagement between 2016 and 2018.

We argue that our two rounds of surveys appear well positioned to capture something proxying a baseline for outcomes in February 2016, before the project was disbursing substantially, and improvements over a two year period as the project is actively implemented. The 2018 Quality Assurance Review (a report that reviews the project's implementation and is led by an external consultant) indicates that disbursements of project funds were delayed in the early years of the project, with funds starting to flow and investment activities happening only towards the middle of 2017, i.e, the end of the World Bank's fiscal year 2016/17 (Roelcke et al., 2018). Our survey measures are consistent with this. For example, on a set of questions posed to local government officials about whether disbursements from the central government are timely, responses show significant improvement in April 2018, compared to February 2016, consistent with the consultants' report. However, there is no difference in this improvement between the project and non-project LGAs.

Third, measuring experience with service delivery is subject to error and reporting biases. For example, the period 2016-2018 in Tanzania is one where a new president took office (in October 2015), announced and implemented several policy measures to crack down on corruption, and strengthened performance norms among government officials. These announcements and actions may have created a perception that things are improving, coloring the survey responses equally in both project and non-project areas. However, our survey is careful to avoid pure perception questions such as degree of "satisfaction" with government performance or agreement/disagreement with whether things are "improving." Instead, the questions probe for actual

experience, and some show improvements over time only in the non-project LGAs, not in project LGAs. For example, in 2016, only 29 percent of government officials reported paying property taxes on commercial building they own, which rose to 62 percent in 2018, compared to around 75 percent in both years in the project LGAs.

Fourth, despite our checks of pre-project parallel trends in a key measure of local state capacity—own source revenues, since the project LGAs were not selected at random, there remains a possibility that their trends would have been different in the absence of the program, leading to bias in our difference-in-difference estimates. Nevertheless, the following facts make it difficult to defend an argument that the project had impact that we are unable to discern: that the survey data are exceptionally rich, gathering data on many different aspects of local state capacity, from the experience of both government officials and citizens; that we find no evidence at all of significant improvements in the project areas that are different from improvements in the non-project areas; and that the improvements reported over time in both project and non-project areas are consistent with country-wide changes in Tanzania, regardless of the project. For example, the new President re-centralized the collection of local property taxes. The re-centralization of local property taxes directly affected the project design. While at the outset, one of the main indicators of increasing local state capacity was expected to be increases in local own source revenues, the APA entirely dropped the scoring of this component in its 6th round (2017-2018). Yet, we find increases in local government official reports of local tax efforts, and especially so in non-project areas. Local officials' payment of their own property taxes increases substantially in non-project areas, and they also are more likely than project-area officials to report efforts to improve local tax collection.

To offer ideas and recommendations from research for further innovation in the design of such projects in Tanzania and beyond, we closely reviewed project documents to understand the general design and theory of change on which these capacity building projects are founded. We find that these projects share a common, global template that is being applied across different countries and contexts, centered on the role of an Annual Performance Assessment which is expected to verify whether local governments have certain institutional features that are found in high state capacity countries: such as, existence of planning documents, council meeting minutes, audit reports, procurement tenders and the like.[26] Qualitative research has critiqued this approach to building state capacity as "isomorphic mimicry" (Andrews et al., 2017), whereby developing countries are made to produce documents and establish protocols that resemble institutions in donor countries, but fail to effectively perform the functions of a state. The lack of difference in measured outcomes of state functioning in the data from Tanzania offers quantitative evidence that is consistent with such critiques.

Alternatively, the amounts committed or the scope of the capacity building component might be too small to make a dent.[27] Further, the rationale for using a financial incentive approach is

[26] See examples from Egypt (World Bank, 2018a), Ethiopia (World Bank, 2014b), India (West Bengal) (World Bank, 2016a), Indonesia (World Bank, 2005), Uganda (World Bank, 2016b), and Vietnam (World Bank, 2014a).

[27] For the project in Tanzania, the total donor contribution was US$255 million, but only US$54 million was

not necessarily substantiated by strong *ex-ante* evidence and hence it cannot be ruled out that a more traditional development financing approach (without explicit incentives) could be more effective, as recent evidence from the health sector demonstrates (Kandpal et al., 2020).

6 Conclusion

Going forward, we recommend, first, that projects targeted at building state capacity invest more resources in learning through policy experimentation within the project, given the lack of established knowledge about how state capacity comes and the existing critiques of donor approaches to transplant formal institutions (Bourguignon & Gunning, 2018; Andrews et al., 2017; World Bank, 2017). Second, we recommend more research to understand forces of change in countries that may be strengthening incentives of local governments to deliver services. From reviews of research available so far, it seems that greater political contestation and demands from citizens for improved governance and service delivery are behind these improving incentives, but nevertheless with several risks and pitfalls (Olken & Pande, 2013; World Bank, 2016c; Dal Bo & Finan, 2016; Bardhan & Mookherjee, 2016; Khemani, 2019). A deeper analysis of administrative data can be an essential asset for both of these recommendations.[28]

Project designs going forward can use the recommendations from these reviews to strengthen political incentives for service delivery, and thus enable the emergence of state capacity along similar lines as how such capacity emerged in today's rich countries (Besley & Persson, 2009; Fukuyama, 2004; World Bank, 2016c; Khemani, 2019). For example, performance assessments could focus on rigorously measuring service delivery performance—e.g., road connectivity, garbage collection, coverage of drainage and sewage systems—rather than on primarily reviewing documents and protocols as is currently being done through the APAs. Project design could focus on the communication of, and deliberation around, these performance assessments, with citizens, especially through mass media whose role in strengthening institutions has been recently emphasized in research (World Bank, 2016c; La Ferrara, 2016). Investing in communication and deliberation is not a soft option but rather one that could be applied more scientifically through dedicated projects aimed at building state capacity,

References

Abadie, A. (2005). Semiparametric Difference-in-Differences Estimators. *The Review of Economic Studies*, *72*(1), 1–19. 67

Acemoglu, D. (2005). Politics and Economics in Weak and Strong States. *Journal of Monetary Economics*, *52*(7), 1199–1226. 1, 4, 6

allocated to the capacity building component, with the rest going to the financing of local infrastructure.

[28] See a recent discussion on this in Legovini & Jones (2020).

Acemoglu, D., García-Jimeno, C., & Robinson, J. A. (2015). State Capacity and Economic Development: A Network Approach. *American Economic Review*, *105*(8), 2364–2409. 1, 4, 6, 7, 13, 25

Acemoglu, D., & Robinson, J. A. (2020). *The Narrow Corridor.*. Penguin Press. OCLC: 1135576051. 7

Anderson, M. L. (2008). Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects. *Journal of the American Statistical Association*, *103*(484), 1481–1495. Publisher: Taylor & Francis. 16, 43

Andrews, M., Pritchett, L., & Woolcock, M. (2013). Escaping Capability Traps Through Problem Driven Iterative Adaptation (PDIA). *World Development*, *51*, 234–244. 5

Andrews, M., Pritchett, L., & Woolcock, M. J. V. (2017). *Building state capability: evidence, analysis, action.* Oxford ; New York, NY: Oxford University Press, first edition ed. OCLC: ocn973093519. 5, 27, 28

Anter, A. (2020). The Modern State and Its Monopoly on Violence. ISBN: 9780190679545. 7

Bandyopadhyay, S., & Green, E. (2016). Precolonial Political Centralization and Contemporary Development in Uganda. *Economic Development and Cultural Change*, *64*(3), 471–508. Publisher: The University of Chicago Press. 1, 5

Banerjee, A. V., & Duflo, E. (2012). *Poor economics: a radical rethinking of the way to fight global poverty*. New York: PublicAffairs, paperback first published ed. OCLC: 798933070. 5

Bardhan, P., & Mookherjee, D. (2006). Pro-poor targeting and accountability of local governments in West Bengal. *Journal of Development Economics*, *79*(2), 303–327. 5

Bardhan, P. K., & Mookherjee, D. (2000). Capture and Governance at Local and National Levels. *American Economic Review*, *90*(2), 135–139. 5

Bardhan, P. K., & Mookherjee, D. (2016). Clientelistic Politics and Economic Development: An Overview. *EDI Working Paper Series*, *WP16/10.111.5*. 5, 28

Battaile, W. G. (2020). What does Tanzania's move to lower-middle income status mean? *Africa Can End Poverty blog*. 5

Besley, T. (2020). State Capacity, Reciprocity, and the Social Contract. *Econometrica*, *88*(4), 1307–1335. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA16863. 4, 6, 13

Besley, T., & Ghatak, M. (2005). Competition and Incentives with Motivated Agents. *American Economic Review*, *95*(3), 616–636. 4

Besley, T., & Persson, T. (2009). The Origins of State Capacity: Property Rights, Taxation, and Politics. *American Economic Review*, *99*(4), 1218–1244. 1, 4, 6, 7, 28

Besley, T., & Persson, T. (2011). Pillars of Prosperity: The Political Economics of Development Clusters. Economics Books, Princeton University Press. 1, 6, 7

Bisin, A. (2020). A Comment on: "State Capacity, Reciprocity, and the Social Contract" by Timothy Besley. *Econometrica*, *88*(4), 1345–1349. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18010. 4, 13

Bloom, N., & Van Reenen, J. (2010). Human Resource Management and Productivity. Tech. Rep. w16019, National Bureau of Economic Research, Cambridge, MA. 13

Bourguignon, F., & Gunning, J. W. (2018). Foreign Aid and Governance: a Survey. In *Handbook of Economic Development and Institutions*. Princeton University Press. 5, 28

Bourguignon, F., & Platteau, J.-P. (2015). The Hard Challenge of Aid Coordination. *World Development*, *69*(C), 86–97. Publisher: Elsevier. 5

Bourguignon, F., & Sundberg, M. (2007). Aid Effectiveness – Opening the Black Box. *American Economic Review*, *97*(2), 316–321. 5

Bowles, S. (2020). A Comment on: "State Capacity, Reciprocity, and the Social Contract" by Timothy Besley. *Econometrica*, *88*(4), 1337–1343. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18037. 4, 13

Brautigam, D. A., & Knack, S. (2004). Foreign Aid, Institutions, and Governance in Sub-Saharan Africa. *Economic Development and Cultural Change*, *52*(2), 255–85. Publisher: University of Chicago Press. 5

Dal Bo, E., & Finan, F. (2016). At the Intersection: A Review of Institutions in Economic Development. 6, 28

Dal Bó, E., Finan, F., & Rossi, M. A. (2013). Strengthening State Capabilities: The Role of Financial Incentives in the Call to Public Service. *The Quarterly Journal of Economics*, *128*(3), 1169–1218. 4, 5, 7

Dal Bó, P., Foster, A., & Putterman, L. (2010). Institutions and behavior: Experimental evidence on the effects of democracy. *American Economic Review*, *100*(5), 2205–29. URL https://www.aeaweb.org/articles?id=10.1257/aer.100.5.2205 6

de Janvry, A., & Dethier, J.-J. (2012). The World Bank and Governance: The Bank's Efforts to Help Developing Countries Build State Capacity. SSRN Scholarly Paper ID 2179406, Social Science Research Network, Rochester, NY. 1, 7

Dell (2010). The Persistent Effects of Peru's Mining Mita. *Econometrica*, *78*(6), 1863–1903. 5

Dell, M., Lane, N., & Querubin, P. (2018). The Historical State, Local Collective Action, and Economic Development in Vietnam. *Econometrica*, *86*(6), 2083–2121. 5

Dell, M., & Olken, B. (2020). The Development Effects of the Extractive Colonial Economy: The Dutch Cultivation System in Java. *Review of Economic Studies*, *87*(1), 164–203. 5

Devarajan, S., Khemani, S., & Shah, S. (2009). The Politics of Partial Decentralization. In *Does decentralization enhance service delivery and poverty reduction?*. Edward Elgar Publishing. 5, 6

Dincecco, M., & Katz, G. (2016). State Capacity and Long-run Economic Performance. *The Economic Journal*, *126*(590), 189–218. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/ecoj.12161. 1

Erman, A. E., Solis Uehara, C. C., & Beaudet, C. (2021). Leveling up : Impacts of performance-based grants on municipal revenue collection in Mozambique. *Policy Research working paper, World Bank*, (9789). 1

Faguet, J.-P. (2003). Does Decentralization Increase Government Responsiveness to Local Needs? Evidence from Bolivia. SSRN Scholarly Paper ID 373600, Social Science Research Network, Rochester, NY. 5

Faguet, J.-P. (2014). Decentralization and Governance. *World Development*, *53*, 2–13. 5

Fergusson, L., Molina, C. A., & Robinson, J. (2020). The Weak State Trap. SSRN Scholarly Paper ID 3651454, Social Science Research Network, Rochester, NY. 1, 4

Ferraz, C., Finan, F., & Martinez-Bravo, M. (2020). Political Power, Elite Control, and Long-Run Development: Evidence from Brazil. Tech. Rep. w27456, National Bureau of Economic Research, Cambridge, MA. 6

Fukuyama, F. (2004). The Imperative of State-Building. *Journal of Democracy*, *15*, 17–31. 6, 28

Gennaioli, N., & Rainer, I. (2007). The Modern Impact of Precolonial Centralization in Africa. *Journal of Economic Growth*, *12*(3), 185–234. Publisher: Springer. 1, 5

Gibson, J., Olivia, S., Boe-Gibson, G., & Li, C. (2021). Which night lights data should we use in economics, and where? *Journal of Development Economics*, *149*, 102602. 19

Government of the United Republic of Tanzania (2013). A Study on Local Government Authorities' Own Source Revenues. 12

Hanson, J. K., & Sigman, R. (2019). State Capacity and World Bank Project Success. *unpublished working paper*. 1

Herbst, J. I. (2000). *States and power in Africa: comparative lessons in authority and control*. Princeton studies in international history and politics. Princeton, N.J: Princeton University Press. 1

Hjort, J., Moreira, D., Rao, G., & Santini, J. F. (2019). How Research Affects Policy: Experimental Evidence from 2,150 Brazilian Municipalities. SSRN Scholarly Paper ID 3406456, Social Science Research Network, Rochester, NY. 5

Independent Evaluation Group (2008). Decentralization in Client Countries: An Evaluation of World Bank Support, 1990–2007. 1, 7

Independent Evaluation Group (2018). Approach paper: Evaluation of the World Bank Group engagement on strengthening subnational governments. 1, 5

Jones, S., Imber, V., Gray, J., Lawson, A., Ghartey, A., Kashangaki, J., Maphiri, D., Mwanwina, I., Therkildsen, O., Wyatt, A., Kotoglou, K., & Shah, A. (2006). Developing capacity: An evaluation of DFID-funded technical cooperation for economic management in Sub-Saharan Africa. 7

Kandpal, E., Khanna, M., Loevinsohn, B., Pradhan, E., Fadeyibi, O., McGee, K., Odutolu, O., Fritsche, G. B., Meribole, E., & Vermeersch, C. (2020). The effect of direct facility financing, autonomy, community engagement, supervision, and performance-based payments in strengthening primary health care: a large scale quasi-experimental trial in Nigeria. 28

Khemani, S. (2015). Political Capture of Decentralization: Vote Buying through Grants to Local Jurisdictions. In *Is Decentralization Good for Development? Perspectives from Academics and Policy Makers*. Oxford University Press. 5

Khemani, S. (2019). What Is State Capacity ? Tech. Rep. 8734, The World Bank. Publication Title: Policy Research Working Paper Series. 5, 6, 13, 28

Khemani, S. (2020). Legitimacy and Trust in the Times of COVID-19. Tech. Rep. Research & Policy brief No. 32, World Bank, Washington DC. 6

Knack, S. (2001). Aid Dependence and the Quality of Governance: Cross-Country Empirical Tests. *Southern Economic Journal*, *68*(2), 310–329. Publisher: Southern Economic Association. 5

La Ferrara, E. (2016). Media as a tool for institutional change in development. 28

Legovini, A., & Jones, M. R. (2020). Administrative data in research at the World Bank: The case of Development Impact Evaluation (DIME). In *Handbook on Using Administrative Data for Research and Evidence-based Policy*. https://admindatahandbook.mit.edu/. 28

Levy, B., & Kpundeh, S. (2004). Building state capacity in Africa: New approaches, emerging lessons. 7

Li, X., Zhou, Y., Zhao, M., & Zhao, X. (2020). A harmonized global nighttime light dataset 1992–2018. *Scientific Data*, *7*(1). 61

Lizzeri, A., & Persico, N. (2004). Why did the Elites Extend the Suffrage? Democracy and the Scope of Government, with an Application to Britain's "Age of Reform". *The Quarterly Journal of Economics*, *119*(2), 707–765. Publisher: Oxford Academic. 4, 7

Lowes, S., Nunn, N., Robinson, J. A., & Weigel, J. (2017). The Evolution of Culture and Institutions: Evidence from the Kuba Kingdom. *Econometrica*, *85*(4), 1065–1091. 5

Martinez-Bravo, M., Mukherjee, P., & Stegmann, A. (2017). The Non-Democratic Roots of Elite Capture: Evidence From Soeharto Mayors in Indonesia. *Econometrica*, *85*(6), 1991–2010. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA14125. 6

Michalopoulos, S., & Papaioannou, E. (2013). Pre-Colonial Ethnic Institutions and Contemporary African Development. *Econometrica*, *81*(1), 113–152. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA9613. 1, 5

Migdal, J. S. (1988). *Strong societies and weak states: state-society relations and state capabilities in the Third World*. Princeton, N.J: Princeton University Press. 1

Muralidharan, K., Niehaus, P., & Sukhtankar, S. (2016). Building State Capacity: Evidence from Biometric Smartcards in India. *American Economic Review*, *106*(10), 2895–2929. 4, 5, 7

Myerson, R. B. (2011). Rethinking the Fundamentals of State-Building. *Working Paper*. 5

Olken, B. (2007). Monitoring Corruption: Evidence from a Field Experiment in Indonesia. *Journal of Political Economy*, *115*(2), 200–249. Publisher: The University of Chicago Press. 11

Olken, B., & Pande, R. (2013). J-PAL Governance Initiative Review Paper. *Abdul Latif Jameel Poverty Action Lab*. 6, 28

Papaioannou, E. (2020). A Comment on: "State Capacity, Reciprocity, and the Social Contract" by Timothy Besley. *Econometrica*, *88*(4), 1351–1358. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA18028. 4, 13

Pritchett, L., Woolcock, M., & Andrews, M. (2013). Looking Like a State: Techniques of Persistent Failure in State Capability for Implementation. *The Journal of Development Studies*, *49*(1), 1–18. Publisher: Routledge _eprint: https://doi.org/10.1080/00220388.2012.709614. 5

Rajan, R., & Subramanian, A. (2007). Does Aid Affect Governance? *American Economic Review*, *97*(2), 322–327. 5, 7

Rajan, R., & Subramanian, A. (2008). Aid and Growth: What Does the Cross-Country Evidence Really Show? *The Review of Economics and Statistics*, *90*(4), 643–665. Publisher: MIT Press. 5, 7

Rasul, I., & Rogger, D. (2018). Management of Bureaucrats and Public Service Delivery: Evidence from the Nigerian Civil Service. *The Economic Journal*, *128*(608), 413–446. 13

Roelcke, G., Nkya, F. M., Besta, N. A., Mchallo, I., Musyangi, M. E., & Orao, S. (2018). Urban local government strengthening program united republic of tanzania (program-for-results operation): Quality assurance review of the sixth annual performance assessment (assessing the performance during fy 2016/17). 26

Stern, N. H., Dethier, J.-J., & Rogers, F. H. (2005). *Growth and empowerment: making development happen*. Cambridge, Mass.: MIT Press. OCLC: 62034045. 1

Stewart, D. (1795). *Account of the Life and Writings of Adam Smith, LL.D.*. 1

Tilly, C. (1992). *Coercion, capital, and European states, AD 990-1992*. Studies in social discontinuity. Cambridge, MA: Blackwell, rev. pbk. ed ed. 4, 7

Weber, M. (1946). *From Max Weber: Essays in sociology*. New York: Oxford university press. Open Library ID: OL6498314M. 7

Weigel, J. L. (2020). The Participation Dividend of Taxation: How Citizens in Congo Engage More with the State When it Tries to Tax Them. *The Quarterly Journal of Economics*, *135*(4), 1849–1903. Publisher: Oxford Academic. 4

World Bank (2005). Project appraisal document on a proposed loan in the amount of us$14.50 million and a proposed credit in the amount of sdr 9.92 million to the republic of Indonesia for the initiatives for local governance reform project. 27

World Bank (2011). *Conflict, security and development*. No. 33.2011 in World Development Report. Washington, DC: World Bank. OCLC: 730038754. 5

World Bank (2012). Program Appraisal Document on a Proposed Credit in the Amount of SDR 167.6 million (US$255 million equivalent) to the United Republic of Tanzania for an Urban Local Government Strengthening Program (Program-For-Results Operation). 2, 4

World Bank (2014a). International development association program appraisal document on a proposed credit in the amount of SDR 161.80 million (US$250 million equivalent) to the Socialist Republic of Vietnam for the results-based national urban development program in the Northern Mountains Region. 27

World Bank (2014b). International development association program appraisal document on a proposed credit in the amount SDR 245.6 million (US$380 million equivalent) to the Federal Democratic Republic of Ethiopia for the second urban local government development program. 27

World Bank (2016a). Implementation completion and results report (ida-47580) on a credit in the amount of sdr 131.8 million (us$200.0 million equivalent) to the Republic of India for a West Bengal institutional strengthening of Gram Panchayats project. 7, 27

World Bank (2016b). International development association program paper on a proposed additional credit in the amount of SDR 231.70 million (US$335 million equivalent) including US$25 million equivalent from the IDA 18 host community and refugee sub-window and an additional grant from the IDA-18 host community and refugee sub-window in the amount of SDR 17.30 million (US$25 million equivalent) to the Republic of Uganda for the Uganda support to municipal infrastructure development program. 27

World Bank (2016c). Making politics work for development : harnessing transparency and citizen engagement. Tech. Rep. 106337, The World Bank. 5, 6, 28

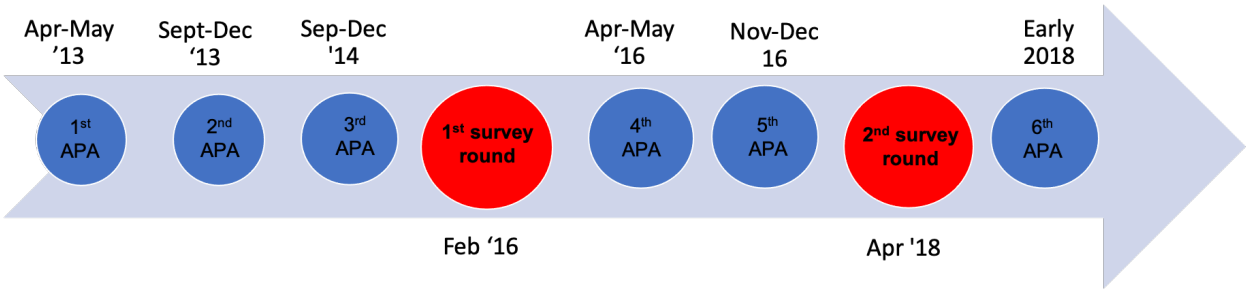
World Bank (Ed.) (2017). *Governance and the law*. No. 2017 in World Development Report. Washington, DC, USA: World Bank Group. OCLC: 976427687. 5, 6, 28

World Bank (2018a). Program appraisal document on a proposed loan in the amount of US$500 million to the Arab Republic of Egypt for an Upper Egypt Local Development Program-For-Results. 27

World Bank (2018b). Quality Assurance Review of the Sixth Annual Performance Assessment, Urban Local Government Strengthening Program, United Republic of Tanzania. 4

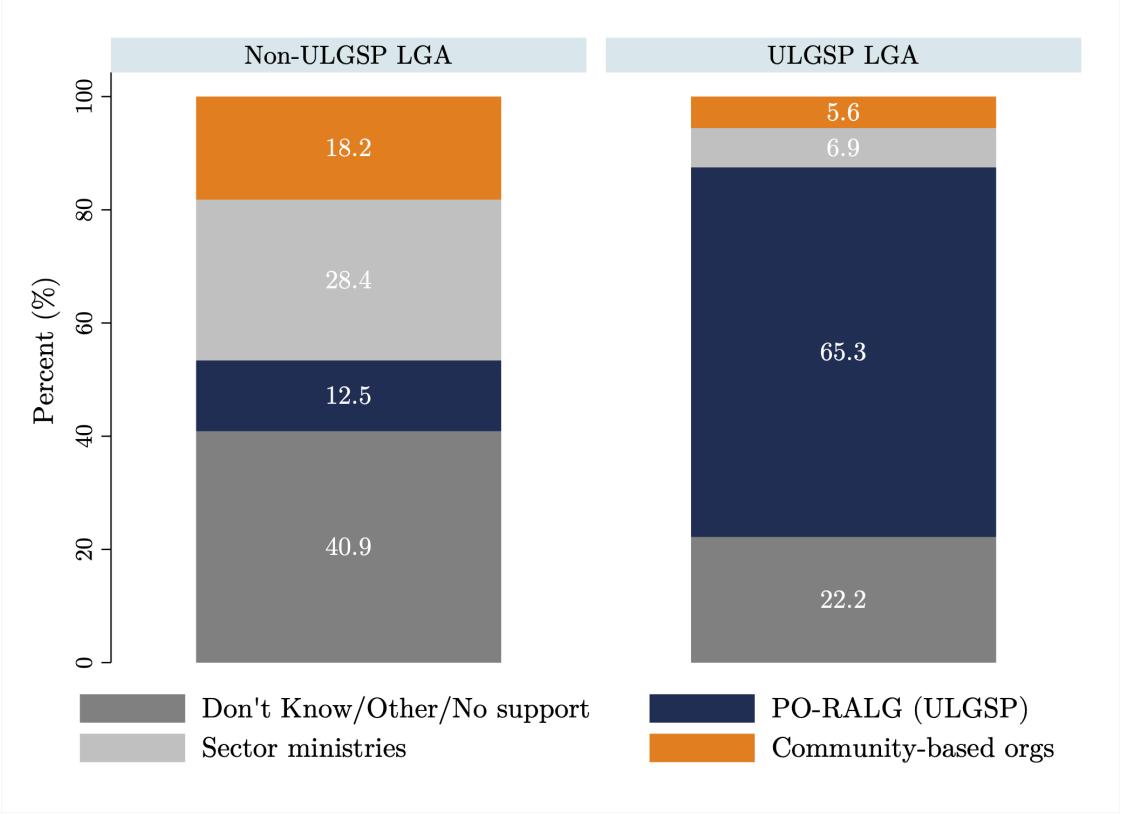
7 Figures and Tables

Figure 1: Timeline of APAs and surveys

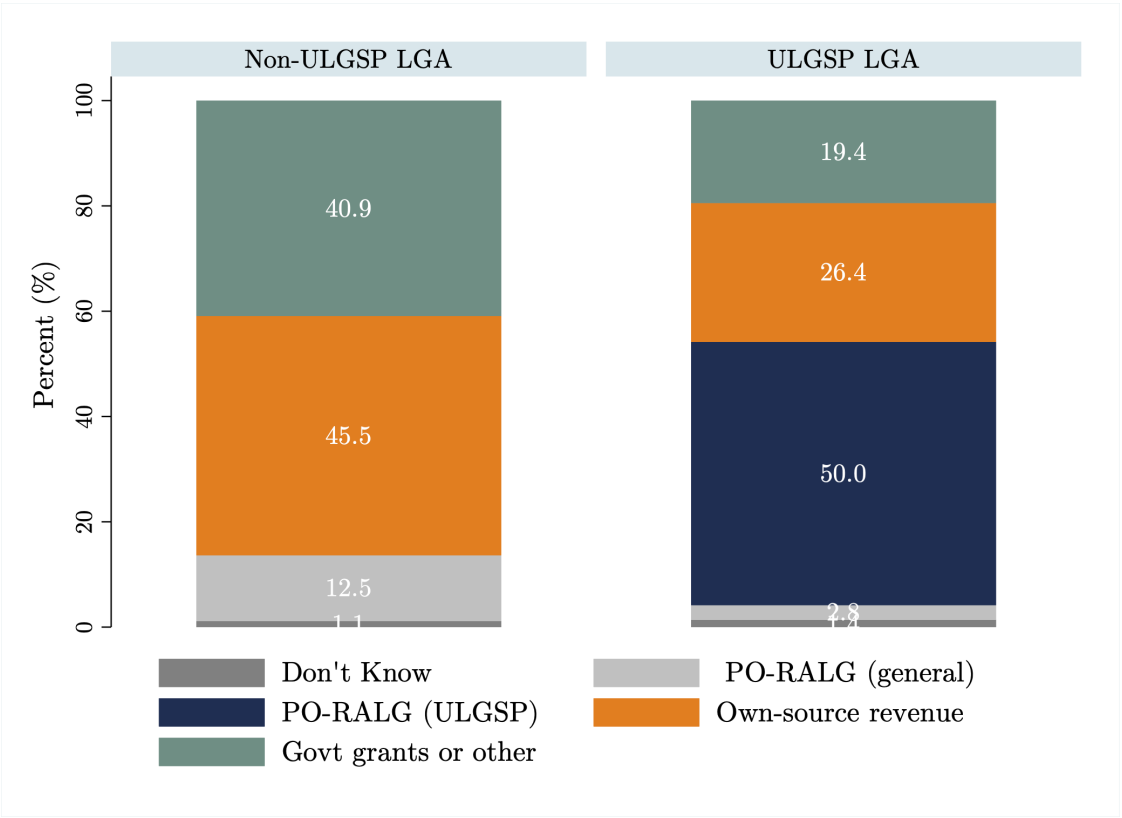


Note: This figure presents the timeline of each one of the Annual Performance Assessments (APAs) and the two rounds of survey data collected for this evaluation.

Figure 2: Awareness of ULGSP support among local officials



(a) Capacity building provision



(b) Infrastructure financing provision

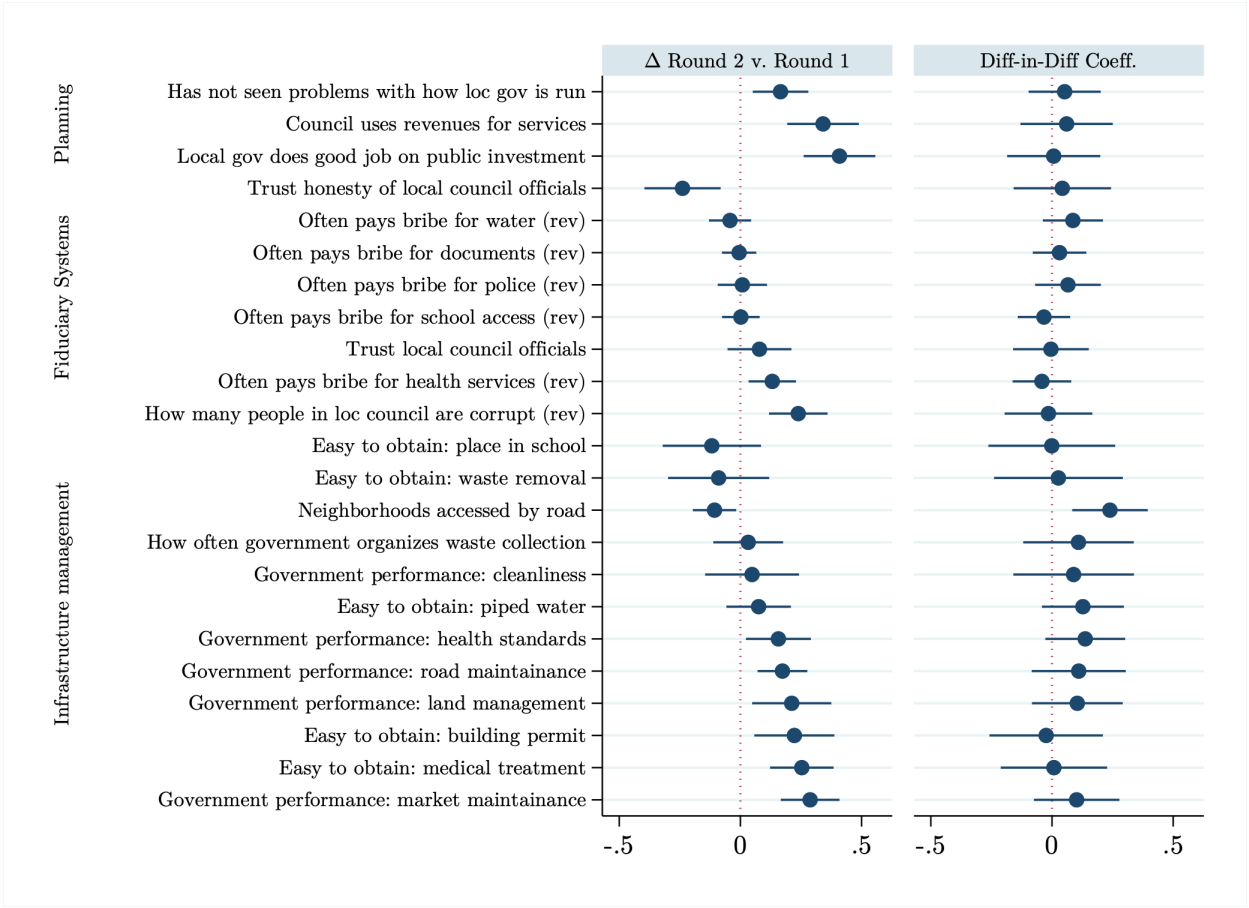


Figure 3: Diff-in-Diff regression results - Households I

Note: This figure reports the coefficients of a difference-in-differences regression including comparators. Each line refers to a regression using the indicated variable as dependent variable. The left panel reports point estimate and 95% CI for the coefficient on the second round survey indicator, while the right panel reports point estimate and 95% CI for the coefficient on the interaction of the second round survey and treatment (project, or ULGSP) indicator, i.e., the differential change in project LGAs between first and second survey rounds. Regressions are performed at the individual level and standard errors are clustered at LGA level.

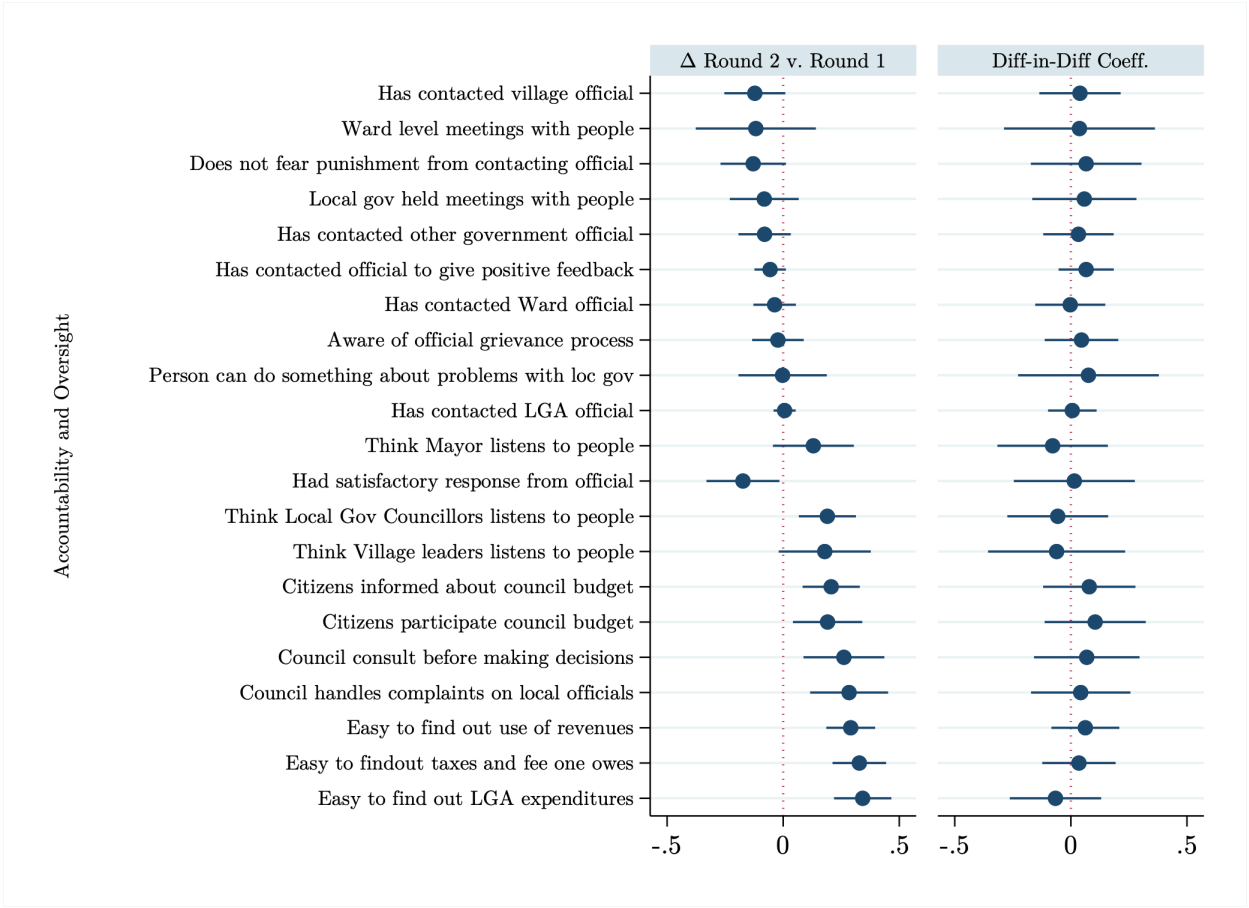


Figure 4: Diff-in-Diff regression results - Households II

Note: This figure reports the coefficients of a difference-in-differences regression including comparators. Each line refers to a regression using the indicated variable as dependent variable. The left panel reports point estimate and 95% CI for the coefficient on the second round indicator, while the right panel reports point estimate and 95% CI for the coefficient on the interaction of the second round and treatment (ULGSP) indicator, i.e., the differential change in ULGSP LGAs between first and second survey rounds. Regressions are performed at the individual level and standard errors are clustered at LGA level.

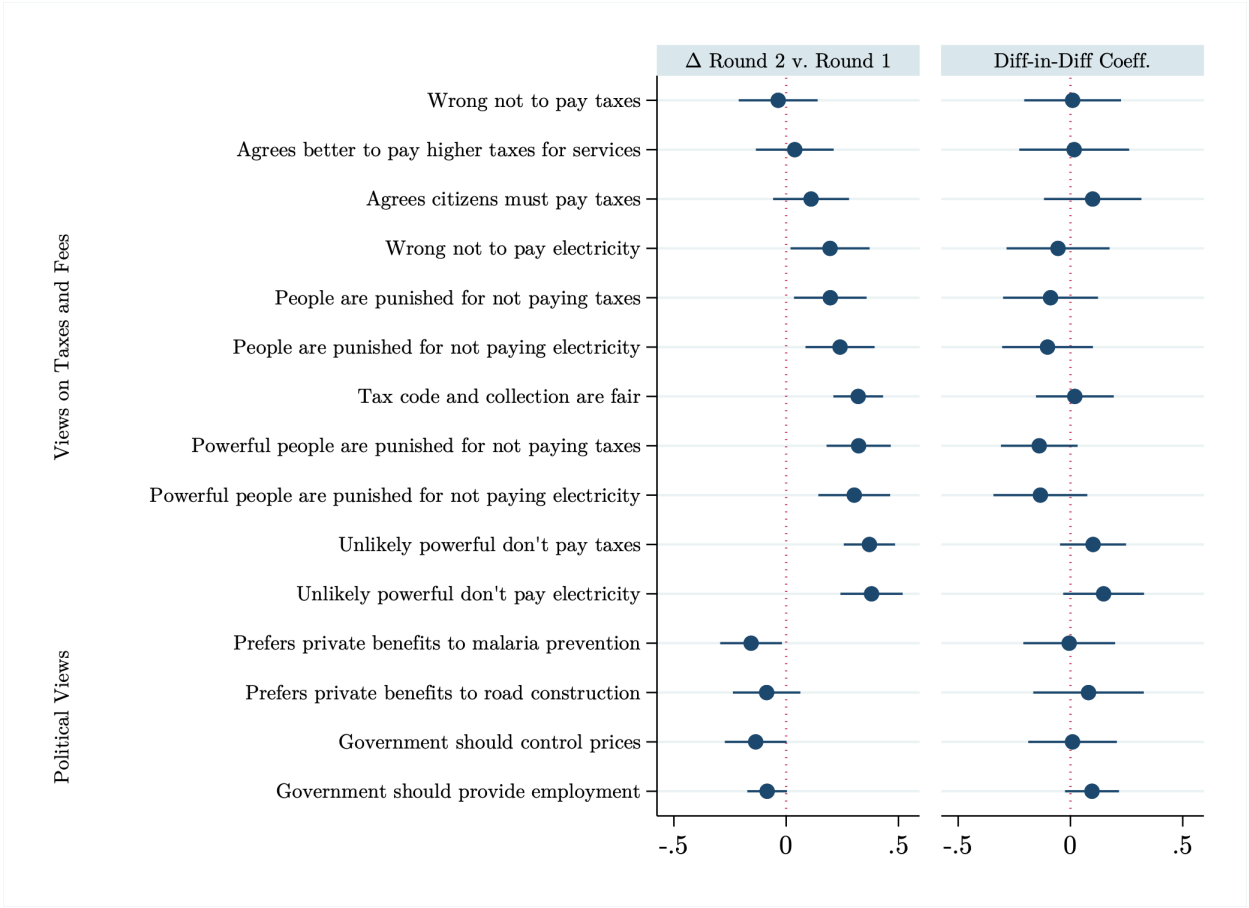


Figure 5: Diff-in-Diff regression results - Households III

Note: This figure reports the coefficients of a difference-in-differences regression including comparators. Each line refers to a regression using the indicated variable as dependent variable. The left panel reports point estimate and 95% CI for the coefficient on the second round indicator, while the right panel reports point estimate and 95% CI for the coefficient on the interaction of the second round and treatment (ULGSP) indicator, i.e., the differential change in ULGSP LGAs between first and second survey rounds. Regressions are performed at the individual level and standard errors are clustered at LGA level.

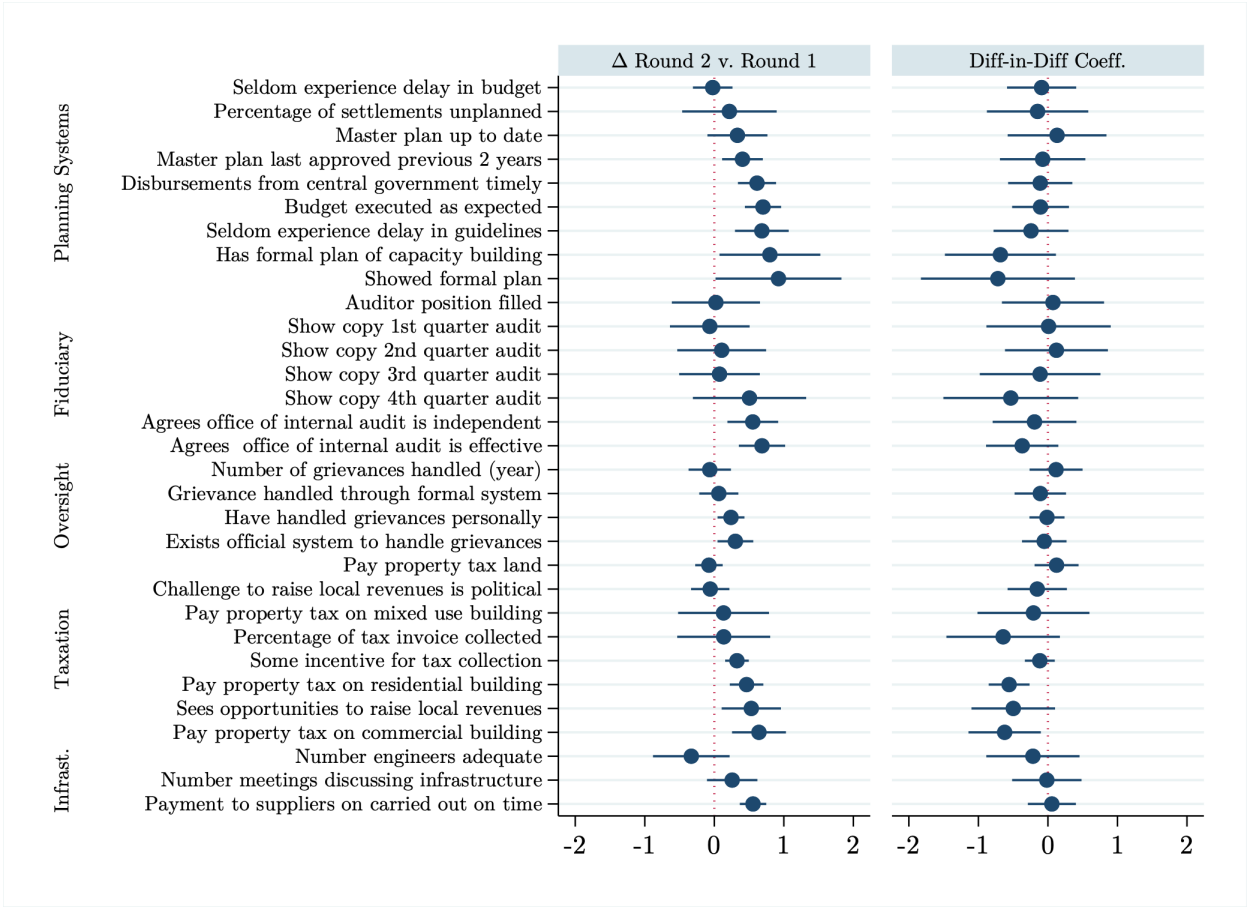


Figure 6: Diff-in-Diff regression results - Government Official I

Note: This figure reports the coefficients of a difference-in-differences regression including comparators. Each line refers to a regression using the indicated variable as dependent variable. The left panel reports point estimate and 95% CI for the coefficient on the second round indicator, while the right panel reports point estimate and 95% CI for the coefficient on the interaction of the second round and treatment (ULGSP) indicator, i.e., the differential change in ULGSP LGAs between first and second survey rounds. Regressions are performed at the individual level and standard errors are clustered at LGA level.

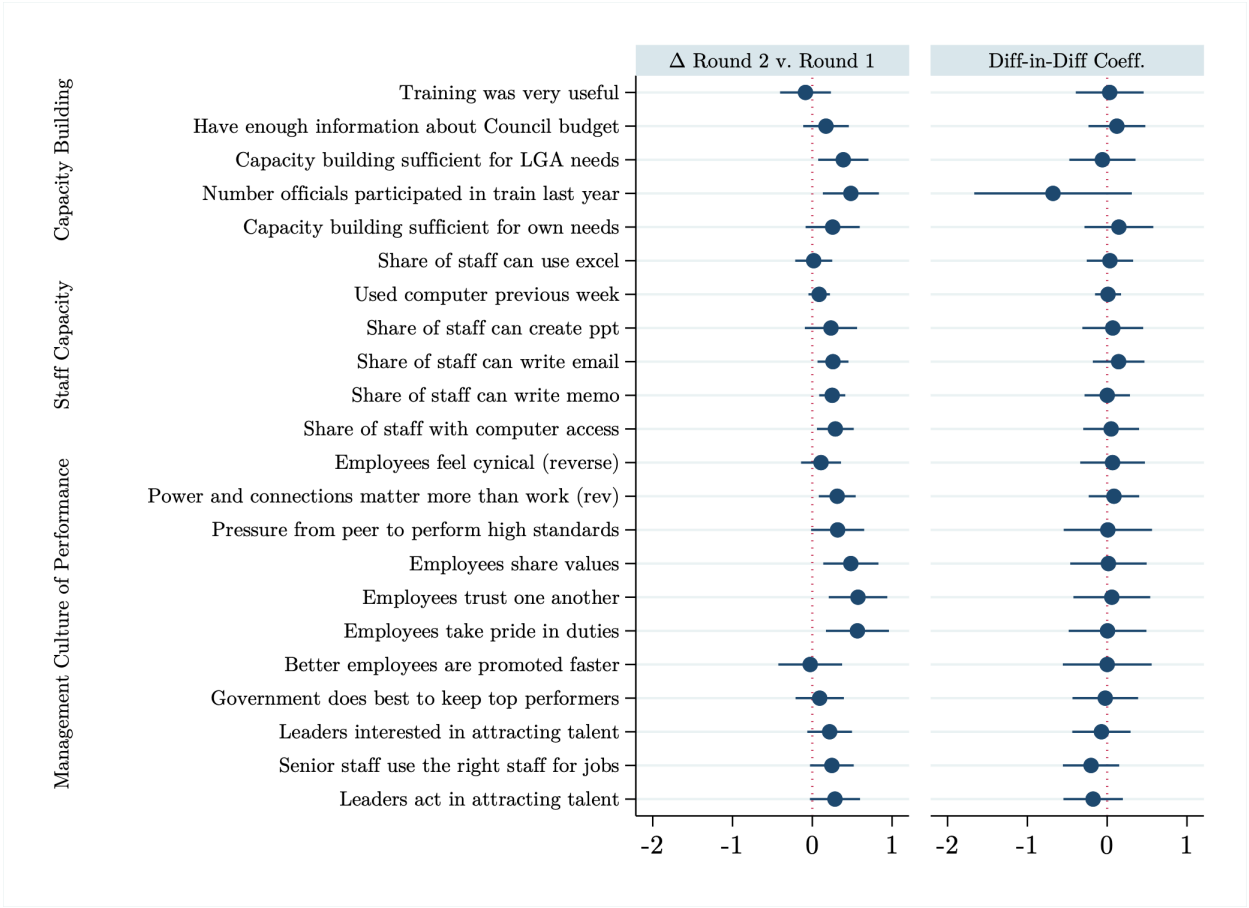
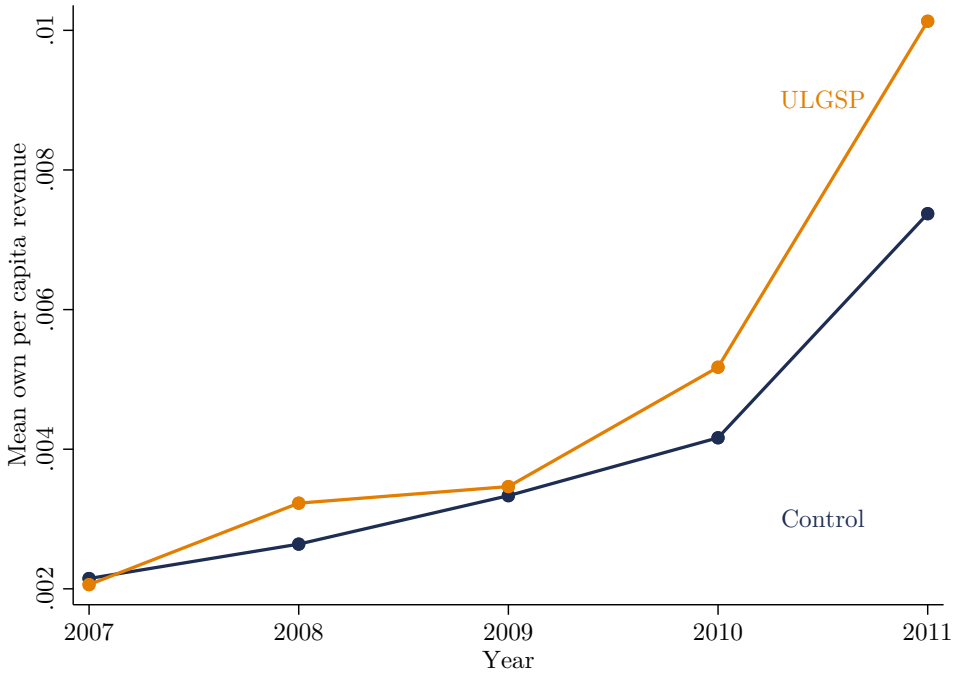


Figure 7: Diff-in-Diff regression results - Government Official II

Note: This figure reports the coefficients of a difference-in-differences regression including comparators. Each line refers to a regression using the indicated variable as dependent variable. The left panel reports point estimate and 95% CI for the coefficient on the second round indicator, while the right panel reports point estimate and 95% CI for the coefficient on the interaction of the second round and treatment (ULGSP) indicator, i.e., the differential change in ULGSP LGAs between first and second survey rounds. Regressions are performed at the individual level and standard errors are clustered at LGA level.

Figure 8: Pre-trends in proxies for state capacity



(a) Own revenue collection

Note: This figure presents trends in own revenue per capita for ULGSP and comparison districts, between 2007 and 2011.

Table 1: Bivariate correlations in household survey

	Urban	Fiduciary	Infrastructure	Accountability	Taxation
Urban	1				
Fiduciary	0.238***	1			
Infrastructure	0.228***	0.237***	1		
Accountability	0.130**	0.108*	0.333***	1	
Taxation	-0.0334	-0.0208	0.296***	0.0157	1

Note: This table presents bivariate correlations between indices created using the inverse-covariance weights for individual items (Anderson, 2008). All indices are created such that higher values correspond to more positive outcomes (e.g. less corruption or better responses of government performance).

Table 2: Round 1 characteristics - Households survey

	(1) ULGSP LGA	(2) Control LGA	(3) Diff	(4) p-value	(5) N
Household size	4.54	5.04	-0.51	0.00	2,998
Literate (all respondents)	0.84	0.75	0.10	0.00	2,998
Education: None	0.10	0.16	-0.06	0.00	2,998
Education:Less than complete Primary	0.13	0.20	-0.06	0.00	2,998
Education:Less than complete lower Secondary	0.50	0.53	-0.03	0.07	2,998
Education:Less than complete upper Secondary	0.23	0.10	0.12	0.00	2,998
Education:College degree	0.04	0.01	0.03	0.00	2,998
Sector of Work: Agriculture	0.47	0.76	-0.29	0.00	2,743
Sector of Work: Small Business	0.32	0.15	0.16	0.00	2,743
Sector of Work: Other	0.22	0.09	0.13	0.00	2,743
Owns a car	0.11	0.03	0.08	0.00	2,998
Owns TV	0.45	0.18	0.27	0.00	2,998
Owns computer	0.13	0.02	0.11	0.00	2,998
Owns video	0.39	0.15	0.24	0.00	2,998
Owns refrigerator	0.25	0.06	0.18	0.00	2,998
Owns stove	0.21	0.05	0.16	0.00	2,998
Owns radio	0.65	0.57	0.08	0.00	2,998
Owns phone	0.88	0.80	0.08	0.00	2,998
Asset Index (ICW)	0.25	-0.20	0.44	0.00	2,998
Government maintains roads well	0.62	0.47	0.15	0.00	2,987
Government maintains prices well	0.52	0.44	0.08	0.00	2,941
Government maintains health standards well	0.67	0.63	0.04	0.04	2,918
Government keeps community clean	0.70	0.66	0.04	0.04	2,932
Government manages use of land well	0.68	0.64	0.04	0.05	2,822
Easy access to building/business permit	0.41	0.34	0.08	0.00	1,661
Easy access to hhd services (water)	0.50	0.38	0.12	0.00	2,186
Easy access to waste collection	0.54	0.47	0.06	0.00	2,211
Easy access to place in primary school	0.93	0.93	0.01	0.37	2,728
Easy access to medical treatment	0.62	0.60	0.03	0.17	2,841
Most or all neighborhoods accessible by paved road	0.08	0.01	0.07	0.00	2,998
Waste collection at least sometimes	0.59	0.31	0.28	0.00	2,998

Note: This table presents average traits of household in ULGSP and control LGAs, and the difference between the means. P-value is reported for difference of means, clustered at LGA-level.

Table 3: Round 1 characteristics - Government officials survey

	(1) ULGSP LGA	(2) Control LGA	(3) Diff	(4) p-value	(5) N
Master plan up to date	0.46	0.27	0.19	0.03	120
Master plan approved in previous 2 years	0.33	0.12	0.21	0.00	120
LGA has formal plan for capacity building	0.94	0.62	0.33	0.02	39
Showed formal plan	0.76	0.46	0.30	0.09	30
Percentage of unplanned settlements	48.96	48.45	0.51	0.94	59
Seldom experience delay in preparing final budget	0.86	0.86	-0.01	0.92	155
Seldom experience delay in guidelines from central gov. for final budget	0.53	0.56	-0.03	0.71	156
Agrees that budget was executed in accordance with expected results	0.16	0.11	0.05	0.14	464
Agrees that disbursements from central government were timely	0.09	0.04	0.05	0.02	453
Position filled? Municipal/City/District Internal Auditor	0.94	0.79	0.16	0.01	120
Agrees that office of internal audit is independent	0.68	0.58	0.10	0.02	459
Agrees that internal audit is effective in monitoring	0.85	0.70	0.15	0.00	467
Number of engineers is adequate	0.19	0.18	0.01	0.89	80
Agrees that payments to suppliers were carried out on time	0.16	0.12	0.04	0.17	456
Number of meetings discussed infrastructure	2.82	2.98	-0.16	0.71	352
Official system to handle grievances	0.88	0.79	0.09	0.01	474
Have handled grievances personally	0.79	0.76	0.03	0.44	474
Number of grievances (year)	329.08	398.52	-69.44	0.45	366
Grievances handled through formal system	0.92	0.92	0.01	0.76	311

Note: This table presents average traits of government officials in ULGSP and control LGAs, and the difference between the means. P-value is reported for difference of means, clustered at LGA-level.

Table 4: Households - Changes in means across waves I

	Means by groups				Difference in means (t-test)	
	(1) ULGSP Round 1 Mean (se)	(2) ULGSP Round 2 Mean (se)	(3) Control Round 1 Mean (se)	(4) Control Round 2 Mean (se)	(5) ULGSP Diff Diff (p-value)	(6) Control Diff Diff (p-value)
Local Council guarantees good use of revenues	0.44	0.66	0.38	0.57	0.22	0.19
	(0.03)	*(0.03)*	*(0.03)*	*(0.03)*	*(0.00)*	*(0.00)*
Local government make good investment plans	0.50	0.71	0.42	0.64	0.21	0.22
	(0.03)	*(0.03)*	*(0.03)*	*(0.04)*	*(0.00)*	*(0.00)*
Have not seen problems with local government	0.85	0.92	0.87	0.92	0.07	0.05
	(0.02)	*(0.01)*	*(0.02)*	*(0.01)*	*(0.00)*	*(0.01)*
Trust Council officials	0.61	0.64	0.57	0.57	0.03	0.00
	(0.02)	*(0.02)*	*(0.03)*	*(0.02)*	*(0.25)*	*(0.90)*
Councillors honest in handling public money	0.88	0.77	0.85	0.75	-0.11	-0.09
	(0.02)	*(0.02)*	*(0.02)*	*(0.03)*	*(0.00)*	*(0.00)*
None or only some of council is corrupt	0.86	0.90	0.84	0.87	0.04	0.03
	(0.02)	*(0.01)*	*(0.01)*	*(0.01)*	*(0.06)*	*(0.13)*
Never paid bribe for documents	0.97	0.98	0.98	0.98	0.00	0.00
	(0.00)	*(0.00)*	*(0.00)*	*(0.00)*	*(0.53)*	*(0.29)*
Never paid bribe for water/sanitation	0.98	0.99	0.99	0.99	0.00	-0.00
	(0.00)	*(0.00)*	*(0.00)*	*(0.00)*	*(0.75)*	*(0.36)*
Never paid bribe for health	0.92	0.95	0.90	0.93	0.03	0.03
	(0.01)	*(0.01)*	*(0.01)*	*(0.01)*	*(0.01)*	*(0.03)*
Never paid bribe to police	0.94	0.96	0.95	0.95	0.02	-0.00
	(0.01)	*(0.01)*	*(0.01)*	*(0.01)*	*(0.06)*	*(0.89)*
Never paid bribe for education	0.99	0.98	0.98	0.98	-0.00	-0.00
	(0.00)	*(0.00)*	*(0.00)*	*(0.00)*	*(0.68)*	*(0.52)*
Government maintains roads well	0.62	0.76	0.47	0.56	0.13	0.09
	(0.04)	*(0.03)*	*(0.04)*	*(0.03)*	*(0.01)*	*(0.00)*
Government maintains prices well	0.52	0.69	0.44	0.58	0.17	0.14
	(0.04)	*(0.03)*	*(0.03)*	*(0.03)*	*(0.00)*	*(0.00)*
Government maintains health standards well	0.67	0.80	0.63	0.70	0.14	0.07
	(0.02)	*(0.02)*	*(0.03)*	*(0.02)*	*(0.00)*	*(0.01)*
Government keeps community clean	0.70	0.77	0.66	0.67	0.07	0.01
	(0.03)	*(0.02)*	*(0.03)*	*(0.02)*	*(0.06)*	*(0.89)*
Government manages use of land well	0.68	0.82	0.64	0.75	0.14	0.11
	(0.02)	*(0.01)*	*(0.03)*	*(0.02)*	*(0.00)*	*(0.00)*
Easy access to building/business permit	0.41	0.51	0.34	0.41	0.10	0.07
	(0.04)	*(0.03)*	*(0.03)*	*(0.02)*	*(0.03)*	*(0.08)*
Easy access to hhd services (water)	0.50	0.64	0.38	0.41	0.14	0.04
	(0.04)	*(0.04)*	*(0.04)*	*(0.03)*	*(0.00)*	*(0.20)*
Easy access to waste collection	0.54	0.51	0.47	0.39	-0.02	-0.08
	(0.04)	*(0.04)*	*(0.04)*	*(0.03)*	*(0.60)*	*(0.10)*
Easy access to place in primary school	0.93	0.92	0.93	0.92	-0.01	-0.01
	(0.01)	*(0.01)*	*(0.01)*	*(0.01)*	*(0.44)*	*(0.64)*
Easy access to medical treatment	0.62	0.77	0.60	0.74	0.15	0.15
	(0.03)	*(0.02)*	*(0.03)*	*(0.02)*	*(0.00)*	*(0.00)*
Most or all neighborhoods accessible by paved road	0.08	0.10	0.01	0.01	0.02	-0.00
	(0.03)	*(0.03)*	*(0.00)*	*(0.00)*	*(0.06)*	*(0.37)*
Waste collection at least sometimes	0.59	0.65	0.31	0.32	0.06	0.01
	(0.05)	*(0.05)*	*(0.03)*	*(0.04)*	*(0.13)*	*(0.73)*

Note: This table presents group averages and differences in means for households matched across the two waves of the panel data. Columns (1)-(4) present average and s.d. for each variable described for ULGSP LGAs and comparator, for each of the survey waves. Columns (5)-(6) present differences in means for the specific groups and the p-value of a t-test for equality of means (clustering at LGA-level)

Table 5: Households - Changes in means across waves II

	Means by groups				Difference in means (t-test)	
	(1)	(2)	(3)	(4)	(5)	(6)
	ULGSP	ULGSP	Control	Control	ULGSP	Control
	Round 1	Round 2	Round 1	Round 2	Diff	Diff
	Mean (se)	Mean (se)	Mean (se)	Mean (se)	Diff (p-value)	Diff (p-value)
Mayor tries to listen	0.09	0.10	0.05	0.09	0.00	0.04
	(0.02)	*(0.01)*	*(0.01)*	*(0.02)*	*(0.89)*	*(0.03)*
Councillors try to listen	0.30	0.35	0.29	0.36	0.04	0.08
	(0.04)	*(0.03)*	*(0.03)*	*(0.03)*	*(0.34)*	*(0.02)*
Mtaa/village leaders try to listen	0.67	0.71	0.64	0.69	0.04	0.06
	(0.04)	*(0.03)*	*(0.04)*	*(0.03)*	*(0.43)*	*(0.27)*
Local Council provide info on budget	0.33	0.47	0.32	0.43	0.14	0.11
	(0.03)	*(0.03)*	*(0.03)*	*(0.03)*	*(0.00)*	*(0.00)*
Local Council allows participation	0.32	0.46	0.32	0.42	0.14	0.10
	(0.03)	*(0.03)*	*(0.02)*	*(0.03)*	*(0.00)*	*(0.00)*
Local Council consults before decisions	0.39	0.56	0.39	0.53	0.17	0.14
	(0.03)	*(0.03)*	*(0.03)*	*(0.04)*	*(0.00)*	*(0.00)*
Local Council handle complaints	0.38	0.56	0.38	0.54	0.18	0.15
	(0.03)	*(0.03)*	*(0.03)*	*(0.04)*	*(0.00)*	*(0.00)*
Ordinary person can affect local governance	0.26	0.27	0.22	0.24	0.01	0.02
	(0.03)	*(0.03)*	*(0.02)*	*(0.03)*	*(0.86)*	*(0.59)*
Ever contacted Mtaa/village official	0.19	0.17	0.17	0.14	-0.02	-0.04
	(0.02)	*(0.01)*	*(0.02)*	*(0.01)*	*(0.30)*	*(0.12)*
Ever contacted Ward official	0.10	0.09	0.07	0.06	-0.01	-0.01
	(0.01)	*(0.01)*	*(0.01)*	*(0.01)*	*(0.70)*	*(0.63)*
Ever contacted LGA official	0.04	0.04	0.02	0.03	0.00	0.00
	(0.01)	*(0.01)*	*(0.00)*	*(0.00)*	*(0.74)*	*(0.59)*
Ever contacted Other govt official	0.05	0.04	0.04	0.02	-0.01	-0.01
	(0.01)	*(0.01)*	*(0.01)*	*(0.00)*	*(0.39)*	*(0.20)*
Ever contacted official for positive feedback	0.05	0.05	0.06	0.04	0.00	-0.01
	(0.01)	*(0.01)*	*(0.01)*	*(0.01)*	*(0.95)*	*(0.09)*
Satisfactory response from official	0.67	0.59	0.56	0.50	-0.08	-0.06
	(0.03)	*(0.04)*	*(0.03)*	*(0.03)*	*(0.13)*	*(0.14)*
Does not fear punishment from contacting officers	0.94	0.93	0.96	0.94	-0.01	-0.03
	(0.02)	*(0.01)*	*(0.01)*	*(0.01)*	*(0.69)*	*(0.11)*
Easy to find out taxes to be paid	0.13	0.26	0.08	0.16	0.13	0.08
	(0.01)	*(0.02)*	*(0.01)*	*(0.02)*	*(0.00)*	*(0.00)*
Easy to find out use of revenues	0.07	0.17	0.06	0.10	0.09	0.05
	(0.01)	*(0.01)*	*(0.01)*	*(0.01)*	*(0.00)*	*(0.00)*
Easy to find out how LGA spend revenues	0.08	0.13	0.04	0.09	0.06	0.05
	(0.01)	*(0.02)*	*(0.01)*	*(0.01)*	*(0.00)*	*(0.00)*
Aware of grievance process at local council	0.22	0.23	0.16	0.16	0.01	-0.00
	(0.02)	*(0.02)*	*(0.02)*	*(0.02)*	*(0.61)*	*(0.87)*
Any Mtaa/village meeting	0.41	0.41	0.51	0.47	-0.00	-0.04
	(0.05)	*(0.05)*	*(0.04)*	*(0.03)*	*(0.92)*	*(0.34)*
Any Ward meeting	0.29	0.25	0.32	0.27	-0.03	-0.05
	(0.04)	*(0.03)*	*(0.04)*	*(0.03)*	*(0.50)*	*(0.40)*

Note: This table presents group averages and differences in means for households matched across the two waves of the panel data. Columns (1)-(4) present average and s.d. for each variable described for ULGSP LGAs and comparator, for each of the survey waves. Columns (5)-(6) present differences in means for the specific groups and the p-value of a t-test for equality of means (clustering at LGA-level)

Table 6: Households - Changes in means across waves III

	Means by groups				Difference in means (t-test)	
	(1)	(2)	(3)	(4)	(5)	(6)
	ULGSP	ULGSP	Control	Control	ULGSP	Control
	Round 1	Round 2	Round 1	Round 2	Diff	Diff
	Mean (se)	Mean (se)	Mean (se)	Mean (se)	Diff (p-value)	Diff (p-value)
Citizens should pay taxes so local govt develops	0.51	0.62	0.42	0.50	0.11	0.08
	(0.03)	*(0.03)*	*(0.03)*	*(0.03)*	*(0.00)*	*(0.05)*
Better to pay high taxes to get more services	0.53	0.55	0.46	0.50	0.02	0.04
	(0.03)	*(0.03)*	*(0.02)*	*(0.03)*	*(0.66)*	*(0.31)*
Not paying taxes is wrong and punishable	0.52	0.50	0.39	0.40	-0.02	0.02
	(0.02)	*(0.03)*	*(0.03)*	*(0.03)*	*(0.56)*	*(0.69)*
Not paying electricity is wrong and punishable	0.76	0.83	0.64	0.73	0.06	0.09
	(0.03)	*(0.02)*	*(0.03)*	*(0.02)*	*(0.11)*	*(0.03)*
Likely to be punished if not paying taxes	0.79	0.81	0.69	0.76	0.02	0.07
	(0.02)	*(0.03)*	*(0.02)*	*(0.03)*	*(0.50)*	*(0.03)*
Tax code and collection is fair	0.54	0.71	0.48	0.65	0.18	0.16
	(0.02)	*(0.02)*	*(0.02)*	*(0.02)*	*(0.00)*	*(0.00)*
Prefers benefit to malaria (count)	0.69	0.54	0.79	0.64	-0.16	-0.15
	(0.05)	*(0.05)*	*(0.05)*	*(0.05)*	*(0.03)*	*(0.03)*
Prefers benefit to road (count)	0.69	0.65	0.74	0.65	-0.04	-0.10
	(0.06)	*(0.06)*	*(0.06)*	*(0.05)*	*(0.67)*	*(0.15)*
Governments should control prices	0.94	0.91	0.94	0.90	-0.03	-0.03
	(0.01)	*(0.01)*	*(0.02)*	*(0.01)*	*(0.11)*	*(0.09)*
Governments should provide employment	0.95	0.96	0.96	0.95	0.00	-0.02
	(0.01)	*(0.01)*	*(0.01)*	*(0.01)*	*(0.67)*	*(0.08)*

Note: This table presents group averages and differences in means for households matched across the two waves of the panel data. Columns (1)-(4) present average and s.d. for each variable described for ULGSP LGAs and comparator, for each of the survey waves. Columns (5)-(6) present differences in means for the specific groups and the p-value of a t-test for equality of means (clustering at LGA-level)

Table 7: Government Officials - Changes in means across waves I

	Means by groups				Difference in means (t-test)	
	(1)	(2)	(3)	(4)	(5)	(6)
	ULGSP	ULGSP	Control	Control	ULGSP	Control
	Round 1	Round 2	Round 1	Round 2	Diff	Diff
	Mean (se)	Mean (se)	Mean (se)	Mean (se)	Diff (p-value)	Diff (p-value)
Master plan up to date	0.46	0.69	0.27	0.44	0.22	0.17
	(0.09)	*(0.09)*	*(0.06)*	*(0.08)*	*(0.13)*	*(0.10)*
Master plan approved in previous 2 years	0.33	0.56	0.12	0.38	0.22	0.26
	(0.08)	*(0.10)*	*(0.04)*	*(0.08)*	*(0.12)*	*(0.00)*
LGA has formal plan for capacity building	0.94	1.00	0.62	0.91	0.06	0.29
	(0.06)	*(0.00)*	*(0.11)*	*(0.06)*	*(0.34)*	*(0.01)*
Showed formal plan	0.76	0.89	0.46	0.85	0.12	0.39
	(0.11)	*(0.08)*	*(0.14)*	*(0.08)*	*(0.40)*	*(0.05)*
Percentage of unplanned settlements	48.96	49.58	48.45	52.74	0.61	4.28
	(3.55)	*(5.15)*	*(4.94)*	*(5.02)*	*(0.89)*	*(0.59)*
Seldom experience delay in preparing final budget	0.86	0.86	0.86	0.92	0.01	0.06
	(0.04)	*(0.04)*	*(0.04)*	*(0.03)*	*(0.92)*	*(0.26)*
Seldom experience delay in guidelines from central gov. for final budget	0.53	0.79	0.56	0.82	0.26	0.26
	(0.09)	*(0.05)*	*(0.07)*	*(0.05)*	*(0.01)*	*(0.01)*
Agrees that budget was executed in accordance with expected results	0.16	0.33	0.11	0.29	0.18	0.18
	(0.04)	*(0.05)*	*(0.02)*	*(0.06)*	*(0.01)*	*(0.01)*
Agrees that disbursements from central government were timely	0.09	0.11	0.04	0.11	0.02	0.07
	(0.03)	*(0.03)*	*(0.01)*	*(0.02)*	*(0.68)*	*(0.01)*
Position filled? Municipal/City/District Internal Auditor	0.94	0.97	0.79	0.81	0.03	0.03
	(0.06)	*(0.03)*	*(0.07)*	*(0.06)*	*(0.67)*	*(0.79)*
Show copy 1st quarter audit	0.94	0.94	0.95	0.95	0.00	0.00
	(0.06)	*(0.06)*	*(0.05)*	*(0.05)*	*(1.00)*	*(1.00)*
Show copy 2nd quarter audit	0.94	1.00	0.91	0.95	0.06	0.05
	(0.06)	*(0.00)*	*(0.06)*	*(0.05)*	*(0.34)*	*(0.58)*
Show copy 3rd quarter audit	0.94	0.94	0.91	0.95	0.00	0.05
	(0.06)	*(0.06)*	*(0.06)*	*(0.05)*	*(1.00)*	*(0.58)*
Show copy 4th quarter audit	0.94	0.94	0.82	0.95	0.00	0.14
	(0.06)	*(0.06)*	*(0.08)*	*(0.05)*	*(1.00)*	*(0.19)*
Agrees that office of internal audit is independent	0.68	0.82	0.58	0.82	0.14	0.24
	(0.08)	*(0.06)*	*(0.07)*	*(0.03)*	*(0.22)*	*(0.00)*
Agrees that internal audit is effective in monitoring	0.85	0.94	0.70	0.93	0.09	0.22
	(0.06)	*(0.02)*	*(0.05)*	*(0.02)*	*(0.19)*	*(0.00)*
Number of engineers is adequate	0.19	0.03	0.18	0.11	-0.17	-0.07
	(0.06)	*(0.03)*	*(0.07)*	*(0.06)*	*(0.01)*	*(0.51)*
Agrees that payments to suppliers were carried out on time	0.16	0.32	0.12	0.25	0.16	0.13
	(0.04)	*(0.05)*	*(0.03)*	*(0.03)*	*(0.01)*	*(0.01)*
Number of meetings discussed infrastructure	2.82	3.76	2.98	4.01	0.94	1.03
	(0.50)	*(0.32)*	*(0.28)*	*(0.43)*	*(0.17)*	*(0.11)*
Official system to handle grievances	0.88	0.96	0.79	0.89	0.08	0.10
	(0.03)	*(0.01)*	*(0.04)*	*(0.03)*	*(0.02)*	*(0.03)*
Have handled grievances personally	0.79	0.88	0.76	0.85	0.09	0.09
	(0.03)	*(0.02)*	*(0.04)*	*(0.02)*	*(0.01)*	*(0.02)*
Number of grievances (year)	329.08	368.47	398.52	350.03	39.39	-48.48
	(48.76)	*(69.37)*	*(119.37)*	*(45.08)*	*(0.62)*	*(0.66)*
Grievances handled through formal system	0.92	0.91	0.92	0.93	-0.01	0.02
	(0.02)	*(0.02)*	*(0.03)*	*(0.02)*	*(0.77)*	*(0.64)*

Note: This table presents group averages and differences in means for GOs matched across the two waves of the panel data. Columns (1)-(4) present average and s.d. for each variable described for ULGSP LGAs and comparator, for each of the survey waves. Columns (5)-(6) present differences in means for the specific groups and the p-value of a t-test for equality of means (clustering at LGA-level).

Table 8: Government Officials - Changes in means across waves II

	Means by groups				Difference in means (t-test)	
	(1)	(2)	(3)	(4)	(5)	(6)
	ULGSP	ULGSP	Control	Control	ULGSP	Control
	Round 1	Round 2	Round 1	Round 2	Diff	Diff
	Mean (se)	Mean (se)	Mean (se)	Mean (se)	Diff (p-value)	Diff (p-value)
Some incentive for tax collection	0.22	0.32	0.09	0.25	0.10	0.16
	(0.02)	*(0.01)*	*(0.03)*	*(0.02)*	*(0.00)*	*(0.00)*
Pay property tax on residential building	0.82	0.82	0.31	0.57	-0.00	0.25
	(0.04)	*(0.03)*	*(0.04)*	*(0.06)*	*(0.94)*	*(0.00)*
Pay property tax on commercial building	0.75	0.76	0.29	0.62	0.00	0.33
	(0.07)	*(0.06)*	*(0.07)*	*(0.08)*	*(0.96)*	*(0.00)*
Pay property tax on mixed use building	0.82	0.88	0.50	0.61	0.06	0.11
	(0.09)	*(0.09)*	*(0.09)*	*(0.13)*	*(0.68)*	*(0.48)*
Pay tax on land	0.49	0.53	0.41	0.39	0.04	-0.02
	(0.05)	*(0.05)*	*(0.04)*	*(0.05)*	*(0.54)*	*(0.73)*
Percentage of tax invoice collected (FY14/15)	67.60	52.80	34.73	39.85	-14.80	5.12
	(4.47)	*(6.23)*	*(7.27)*	*(6.11)*	*(0.06)*	*(0.62)*
Agrees that there are opportunities for raising local revenue	0.85	0.91	0.69	0.91	0.07	0.22
	(0.07)	*(0.02)*	*(0.08)*	*(0.02)*	*(0.36)*	*(0.01)*
Agrees that challenge to raise local revenue is political	0.76	0.71	0.61	0.62	-0.05	0.01
	(0.06)	*(0.04)*	*(0.05)*	*(0.04)*	*(0.46)*	*(0.91)*
Number LGA officials participating in training previous year	46.12	40.17	14.18	33.23	-5.95	19.05
	(12.50)	*(9.11)*	*(2.86)*	*(5.94)*	*(0.71)*	*(0.00)*
Training very useful	0.98	0.97	0.96	0.95	-0.01	-0.01
	(0.02)	*(0.02)*	*(0.02)*	*(0.02)*	*(0.76)*	*(0.82)*
Capacity building activities sufficient to meet LGA needs	0.20	0.31	0.16	0.29	0.11	0.13
	(0.03)	*(0.04)*	*(0.03)*	*(0.05)*	*(0.04)*	*(0.04)*
Capacity building activities sufficient to meet own needs	0.17	0.34	0.18	0.33	0.17	0.14
	(0.02)	*(0.05)*	*(0.03)*	*(0.05)*	*(0.00)*	*(0.04)*
Have enough information about Council budget	0.67	0.82	0.55	0.64	0.15	0.09
	(0.05)	*(0.04)*	*(0.05)*	*(0.05)*	*(0.01)*	*(0.17)*
Used computer within previous week	0.58	0.64	0.55	0.62	0.06	0.07
	(0.03)	*(0.03)*	*(0.02)*	*(0.03)*	*(0.02)*	*(0.01)*
Share of staff with access to computer	0.52	0.64	0.53	0.64	0.13	0.11
	(0.03)	*(0.04)*	*(0.02)*	*(0.03)*	*(0.03)*	*(0.03)*
Share of staff that can write email	0.62	0.73	0.66	0.72	0.11	0.06
	(0.03)	*(0.02)*	*(0.02)*	*(0.02)*	*(0.01)*	*(0.08)*
Share of staff that can create excel spreadsheet	0.58	0.59	0.61	0.59	0.01	-0.02
	(0.02)	*(0.02)*	*(0.02)*	*(0.03)*	*(0.79)*	*(0.62)*
Share of staff that can write a memo	0.61	0.68	0.64	0.70	0.07	0.06
	(0.02)	*(0.02)*	*(0.02)*	*(0.02)*	*(0.06)*	*(0.06)*
Share of staff that can create PPT presentation	0.35	0.43	0.42	0.47	0.09	0.05
	(0.03)	*(0.03)*	*(0.03)*	*(0.05)*	*(0.06)*	*(0.43)*
Leaders interested in attracting and developing talented people	0.78	0.83	0.75	0.83	0.05	0.08
	(0.03)	*(0.04)*	*(0.03)*	*(0.04)*	*(0.27)*	*(0.15)*
Leaders go about attracting and developing talented people	0.77	0.80	0.73	0.83	0.03	0.10
	(0.04)	*(0.04)*	*(0.04)*	*(0.04)*	*(0.49)*	*(0.13)*
Senior staff try to use right staff for right job	0.89	0.90	0.78	0.87	0.01	0.09
	(0.03)	*(0.03)*	*(0.03)*	*(0.03)*	*(0.75)*	*(0.07)*
Better employee would be promoted faster	0.65	0.66	0.65	0.66	0.01	0.01
	(0.06)	*(0.06)*	*(0.05)*	*(0.07)*	*(0.93)*	*(0.93)*
Government does best to keep top performers	0.73	0.74	0.73	0.75	0.01	0.02
	(0.05)	*(0.05)*	*(0.04)*	*(0.05)*	*(0.88)*	*(0.83)*
LGA has flexibility to respond to different needs of community	0.55	0.74	0.36	0.62	0.19	0.26
	(0.06)	*(0.05)*	*(0.06)*	*(0.06)*	*(0.04)*	*(0.00)*
Employees trust one another	0.43	0.77	0.40	0.71	0.34	0.32
	(0.06)	*(0.03)*	*(0.05)*	*(0.04)*	*(0.00)*	*(0.00)*
Employees share strong set of values	0.59	0.78	0.46	0.70	0.19	0.25
	(0.06)	*(0.04)*	*(0.05)*	*(0.04)*	*(0.01)*	*(0.00)*
Employees take pride in fulfilling duties	0.57	0.80	0.44	0.77	0.23	0.33
	(0.06)	*(0.04)*	*(0.06)*	*(0.04)*	*(0.01)*	*(0.00)*
Employees cynical about making a difference (R)	0.85	0.88	0.82	0.84	0.03	0.02
	(0.03)	*(0.02)*	*(0.03)*	*(0.02)*	*(0.47)*	*(0.51)*
Employees feel pressure from colleagues to perform	0.39	0.56	0.36	0.60	0.17	0.24
	(0.08)	*(0.05)*	*(0.04)*	*(0.05)*	*(0.14)*	*(0.00)*
Employees feel political power/connections matter more than performance (R)	0.83	0.90	0.81	0.85	0.07	0.04
	(0.04)	*(0.02)*	*(0.03)*	*(0.03)*	*(0.07)*	*(0.35)*

10c

Table 9: Households - Diff-in-Diff regressions

	(1) Urban Planning Systems	(2) Fiduciary Responsability	(3) Infrastructure Management	(4) Accountability & Transparencey	(5) Views on Taxation and Fees
			Panel A: Simple differences-in-differences		
ULGSP * Round 2 (DiD)	0.030	0.060	0.135	0.015	0.014
	(0.091)	(0.092)	(0.121)	(0.179)	(0.107)
	[-0.155, 0.218]	[-0.136, 0.252]	[-0.109, 0.370]	[-0.348, 0.385]	[-0.208, 0.231]
ULGSP LGA	0.043	-0.007	0.435***	0.075	0.349***
	(0.086)	(0.080)	(0.131)	(0.128)	(0.102)
	[-0.125, 0.209]	[-0.179, 0.151]	[0.168, 0.699]	[-0.203, 0.348]	[0.134, 0.552]
Survey Round 2	0.408***	0.136**	0.263***	-0.056	0.521***
	(0.071)	(0.066)	(0.089)	(0.110)	(0.092)
	[0.260, 0.556]	[-0.00201, 0.272]	[0.0868, 0.452]	[-0.306, 0.181]	[0.325, 0.716]
Observations	5468	3719	2896	976	4492
R-Squared	0.045	0.007	0.092	0.002	0.099
Mean Dep Var	-0.000	0.000	0.000	0.000	0.000
Number clusters (LGA)	40	40	40	40	40
			Panel B: Differences-in-differences including controls		
ULGSP * Round 2 (DiD)	0.038	0.043	0.098	0.013	-0.002
	(0.092)	(0.087)	(0.121)	(0.180)	(0.104)
	[-0.142, 0.228]	[-0.139, 0.226]	[-0.151, 0.350]	[-0.365, 0.379]	[-0.208, 0.210]
ULGSP LGA	0.036	-0.020	0.351***	0.039	0.283***
	(0.082)	(0.080)	(0.127)	(0.138)	(0.101)
	[-0.122, 0.200]	[-0.194, 0.150]	[0.0811, 0.618]	[-0.250, 0.318]	[0.0725, 0.491]
Survey Round 2	0.393***	0.097	0.247***	-0.090	0.508***
	(0.073)	(0.064)	(0.087)	(0.111)	(0.089)
	[0.240, 0.553]	[-0.0381, 0.236]	[0.0757, 0.426]	[-0.322, 0.138]	[0.324, 0.701]
Observations	5462	3717	2892	976	4487
R-Squared	0.079	0.047	0.164	0.035	0.135
Mean Dep Var	-0.000	-0.000	-0.000	0.000	0.001
Number clusters (LGA)	40	40	40	40	40

Note: This table reports regressions using each of the described indices as dependent variable. Non-reported comparators in the second panel include welfare index, household size, age, gender, marital status and education levels. Standard errors clustered at the LGA level are reported in parentheses (* p<0.1, ** p<0.05, *** p <0.01), while 95% confidence-intervals constructed using wild-bootstrapping are reported in brackets.

Table 10: Government Officials - Diff-in-Diff regressions

	(1) Master Plan Up to date	(2) Internal Audit independent	(3) Capacity to raise local revenue	(4) Staff capacity Index	(5) Management Index	(6) Performance Culture Index
			Panel A: Simple differences-in-differences			
ULGSP * Round 2 (DiD)	0.056	-0.260	-0.522	0.100	-0.144	-0.024
	(0.166)	(0.393)	(0.311)	(0.169)	(0.225)	(0.270)
	[-0.280, 0.390]	[-1.066, 0.563]	[-1.164, 0.0863]	[-0.248, 0.449]	[-0.584, 0.313]	[-0.587, 0.507]
ULGSP Indicator	0.190*	0.302	0.547*	-0.149	0.150	0.039
	(0.104)	(0.291)	(0.284)	(0.138)	(0.146)	(0.175)
	[-0.0231, 0.403]	[-0.296, 0.903]	[-0.0578, 1.157]	[-0.444, 0.145]	[-0.139, 0.449]	[-0.305, 0.376]
Survey Round 2	0.167*	0.730***	0.576**	0.121	0.199	0.436**
	(0.095)	(0.232)	(0.218)	(0.112)	(0.184)	(0.179)
	[-0.0315, 0.364]	[0.247, 1.220]	[0.127, 1.065]	[-0.129, 0.346]	[-0.184, 0.579]	[0.0753, 0.811]
Observations	240	917	940	709	874	921
R-Squared	0.085	0.060	0.058	0.010	0.007	0.045
Mean Dep Var	0.454	3.848	4.178	-0.000	-0.000	0.000
Number clusters (LGA)	40	40	40	40	40	40
			Panel B: Differences-in-differences including controls			
ULGSP * Round 2 (DiD)	0.065	-0.249	-0.521	0.085	-0.146	-0.015
	(0.175)	(0.385)	(0.311)	(0.163)	(0.224)	(0.266)
	[-0.297, 0.431]	[-1.023, 0.528]	[-1.154, 0.122]	[-0.263, 0.434]	[-0.610, 0.320]	[-0.561, 0.505]
ULGSP Indicator	0.154	0.287	0.559*	-0.148	0.133	0.020
	(0.113)	(0.281)	(0.283)	(0.134)	(0.143)	(0.172)
	[-0.0865, 0.393]	[-0.337, 0.874]	[-0.0272, 1.161]	[-0.423, 0.135]	[-0.152, 0.417]	[-0.339, 0.389]
Survey Round 2	0.166	0.715***	0.556**	0.155	0.207	0.418**
	(0.106)	(0.233)	(0.220)	(0.104)	(0.190)	(0.178)
	[-0.0596, 0.396]	[0.226, 1.194]	[0.0745, 1.039]	[-0.0602, 0.365]	[-0.195, 0.593]	[0.0327, 0.786]
Observations	240	917	940	709	874	921
R-Squared	0.131	0.081	0.068	0.237	0.022	0.063
Mean Dep Var	0.454	3.848	4.178	-0.000	-0.000	0.000
Number clusters (LGA)	40	40	40	40	40	40

Note: This table reports regressions using each of the described indices as dependent variable. Non-reported comparators in the second panel include welfare index, household size, age, gender, marital status and education levels. Standard errors clustered at the LGA level are reported in parentheses (* p<0.1, ** p<0.05, *** p <0.01), while 95% confidence-intervals constructed using wild-bootstrapping are reported in brackets.

8 Appendix

A Selection of the Comparison Group

Because of the targeted nature of the program (i.e., the program was designed to focus on the 18 LGAs that were among those with faster rates of urbanization at the time the program started), finding a group of comparable set of LGAs was expected to be a difficult task. Additionally, administrative data prior to the program's inception was incomplete for many potential comparison LGAs; some potential comparison LGAs were newly formed and so no historical administrative data were available at all. Against this backdrop, the research team developed a protocol to select

comparison LGAs, drawing on three sources of information:

1. Propensity score matching exercise. The matching process involved a "5-Nearest Neighbors" matching exercise based on population size in 2011, total revenues in 2011, revenue growth between 2007 and 2011, area in sq km in 2012, and population density. The results of this exercise are reported below in Table A1 and Figure A1, which report the coefficients of the probit regression used to estimate the propensity scores and the distribution of the estimated propensity score, respectively. As shown in Figure A1, common support (i.e., finding untreated LGAs with values of the propensity score similar to that of treated LGAs) is very thin for higher values of the propensity score and some treated LGAs are off support (i.e., the estimation cannot identify a comparison LGA that is close enough, in terms of values of propensity score, to the treated LGAs).

2. Survey of treatment LGAs. A short survey was sent to representatives of local offices of PO-RALG in all 18 treatment LGAs. The survey asked representatives to identify the most similar LGA in their own region/zone (compared to their own), the most similar LGA outside of their region, and then to rank the 3 most similar LGAs within own region/zone.

3. Suggestions of potential comparison LGAs from the PO-RALG central team based on their knowledge

Based on the information above, we prepared 2 lists for each program LGA:

1. Survey list: The three most similar LGAs as reported in the survey: (a) most similar within region, (b) most similar by rank beyond the LGA listed in (a), and (c) the most similar LGA outside of the region. If a similar LGA outside of the region was not provided, we replaced it with the next most similar within the region. If any of these listed LGAs was actually in the program, then we moved down the list to the next most similar not in the program. Similarly, if any of these LGAs were participating in the Tanzania Strategic Cities Program (another major World Bank project), then we moved down the list to the next most similar LGA not in the program.

2. Propensity score list: The 2 LGAs with the closest propensity scores.

Finally, we selected the comparison LGA for each program LGA with the following procedure (with allowed repetition of non-program LGAs among comparators) in three steps:

1. If the top comparison LGA in both lists was the same, then that one was chosen.

2. If Step 1 did not hold, then if there was any overlap for the top 2 comparison LGAs across the 2 lists, then we chose the LGA highest on the survey list for which there is overlap between the survey list and the propensity score list.

3. If Step 2 did not hold, then if the top comparison LGA from either list was already selected, we chose the top one from the other list. Otherwise we chose the top comparison LGA from the survey list.

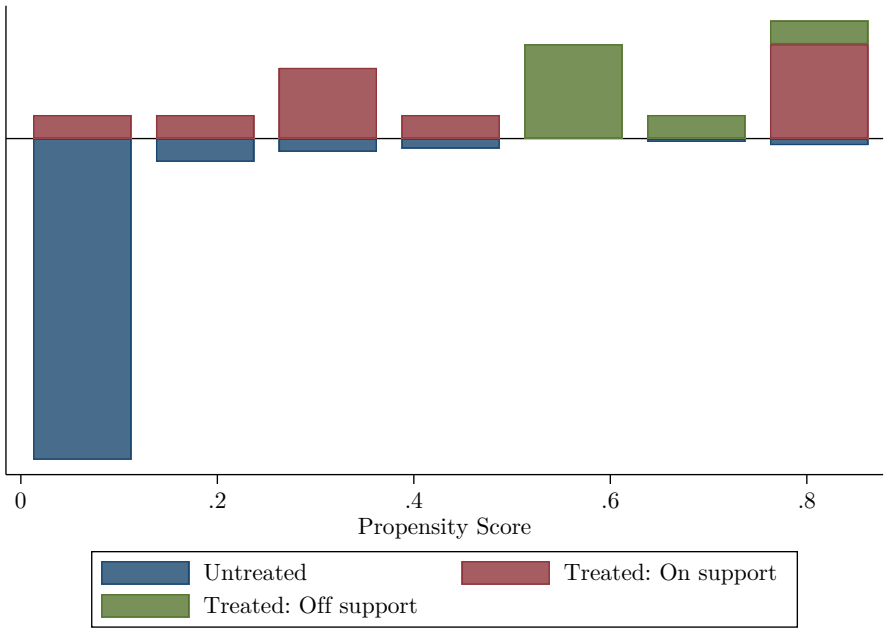
Using the comparison group selected as described above, we find that respondents in project and comparison LGAs are not statistically identical in the main indicators at round 1 (Table 2 and Table 3). This is an unsurprising result, given the targeted nature of the program. Our identification strategy does not rely on project and comparison LGAs being identical but rather that they be developing at similar rates in the absence of the project (often called the "parallel trends" assumption). This is discussed in detail in subsection 3.5.

Table A1: Probit used on propensity-score matching

	(1) program
program	
Population (2011)	-0.00000300
	(-0.79)
Revenue (2011)	-0.0000474
	(-1.11)
Revenue growth (2007-2011)	0.000938
	(0.56)
Area (sqkm)	-0.000638*
	(-2.50)
Population density (2011)	0.000271
	(0.64)
Constant	1.503*
	(2.18)
Observations	130
Pseudo	0.526

Note: This table reports results from a probit regression, where the independent variable is participation in the ULGSP program (* p<0.1, ** p<0.05, *** p <0.01)

Figure A1: Distribution of the propensity score



B Annual Performance Assessment (APA) data and the time path of improvements in project LGAs

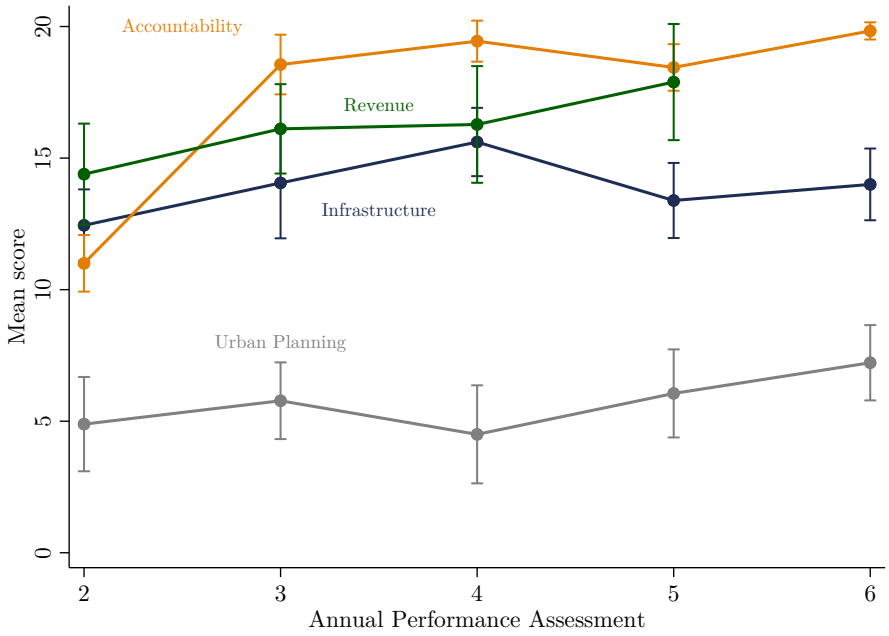
We use information from the Annual Performance Assessment (APA) of the 18 local governments receiving the Urban Local Government Strengthening Program (ULGSP), compiled in the Quality Assurance Reviews (QAR).

We explore two main indicators available throughout the period. Disbursement Linked Indicator (DLI) 2 refers to whether "ULGAs have strengthened institutional performance as scored in the annual performance assessment" and is comprised of five sub-indicators (as described in Section 2 of the paper): Improved urban planning system; increased revenues from property taxes; efficient fiduciary system; improved infrastructure, implementation and O&M; and strengthened accountability and oversight systems. DLI 3 refers to whether "Local infrastructure targets as set out in the annual action plans are met by ULGAs using program funds."

In Figure B1 below we present the mean score for four of the sub-components of DLI2 - we exclude the sub-component on efficient fiduciary system since maximum scores changed over time. As referenced in the timeline presented above, our first round survey was implemented around the same time of the 4th APA (early 2016), and the follow up around the 6th APA (early 2018). The evolution of mean scores is uneven over time. With the exception of the accountability index, that shows a remarkable increase between the 1st and 2nd APAs and then flattens out, the other sub-components are broadly stable over time, with ups and downs and no indication that all improvement happened in the very beginning of the ULGSP program.[29]

[29]We also present the overall changes in APA subcomponents in Table B1

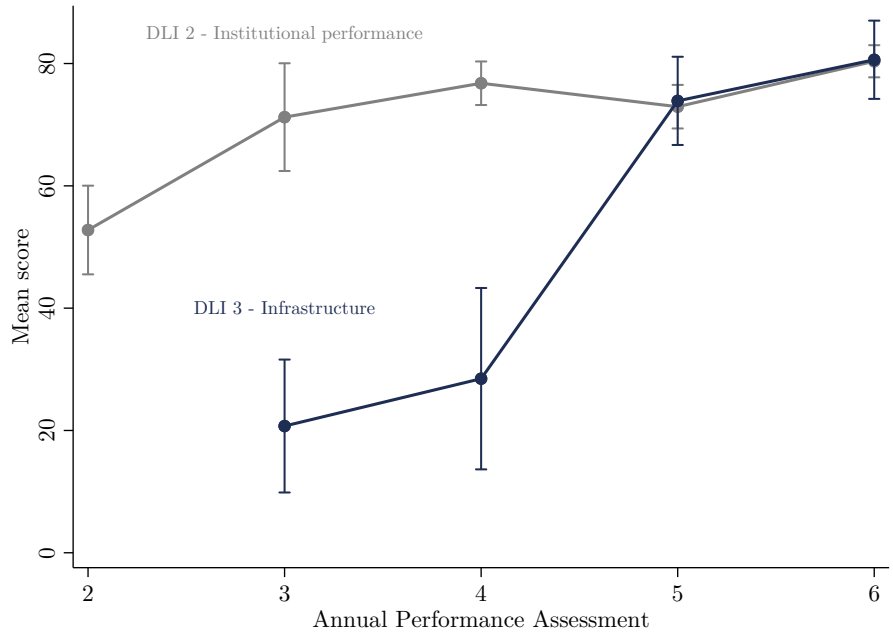
Figure B1: Mean scores on DLI2 sub-components across APAs



Note: This figure presents average scores on each DLI2 component, including 95% CI for the mean. It excludes the "efficient fiduciary system component" since in the 2nd and 3rd APAs the maximum score was lower due to non-inclusion of certain sub-components. The score for "increased revenues from property taxes" was not computed in the 6th APA.

In Figure B2 we present the evolution of average DLI2 and DLI3 components. Focusing first on DLI3, we observe a substantial increase between APAs 4 (early 2016) and 5 (late 2016), with slight improvements in between other APAs. Trends in the DLI2 component are harder to interpret since the maximum score was lower in the 2nd and 3rd APAs (due to lower maximum score on the "efficient fiduciary system" sub-component) and in the 6th APA (due to the absence of "increased revenue from property taxes" sub-component). We use weights to adjust for those, but the composition of the index is not strictly comparable over time.

Figure B2: Mean scores on DLI2 and DLI3 components



Note: This figure presents average scores on DLI2 and DLI3 components, including 95% CI for the mean. It should be noted that DLI2 is not strictly comparable over time: the maximum score for the "efficient fiduciary system" sub-component is lower in 2nd and 3rd APA; and the "increased revenue from property taxes" is absent in the 6th APA, reducing the maximum aggregate score. We use "adjusted scores" provided by the QAR for the 2nd and 3rd APAs, that multiply the final score by 10/9 to expand the maximum score from 90 to 100; and apply the same adjustment ourselves to the 6th APA, multiplying scores by 100/75 to take into account the maximum score of 75 points.

Table B1: Evolution of APA indicators

	Total change	Absolute change		% of total change	
		2-4 APAs	4-6 APAs	2-4 APAs	4-6 APAs
Accountability	8.83	8.44	0.39	0.96	0.04
Improved infrastructure	1.56	3.17	-1.61	2.04	-1.04
Increased revenue	3.50	1.89	1.61	0.54	0.46
Urban planning	2.33	-0.39	2.72	-0.17	1.17

Note: This table reports changes in mean scores for the 18 LGAs receiving the ULGSP program. Column (1) reports the total change in score between the 2nd and 6th APAs, while columns (2) and (3) report changes between 2nd and 4th APAs (before first round survey) and between 4th and 6th APAs (between first and second round surveys), respectively. Columns (5) and (6) express the change in each period as a share of total change. We do not report results for the "Efficient Fiduciary systems" component since the underlying indicators changes throughout the period. The subcomponent "Increased Revenue" was not computed for the 6th APA, so we use the 5th APA as the final round.

Table B2: Evolution of APA indicators vs. survey indicators

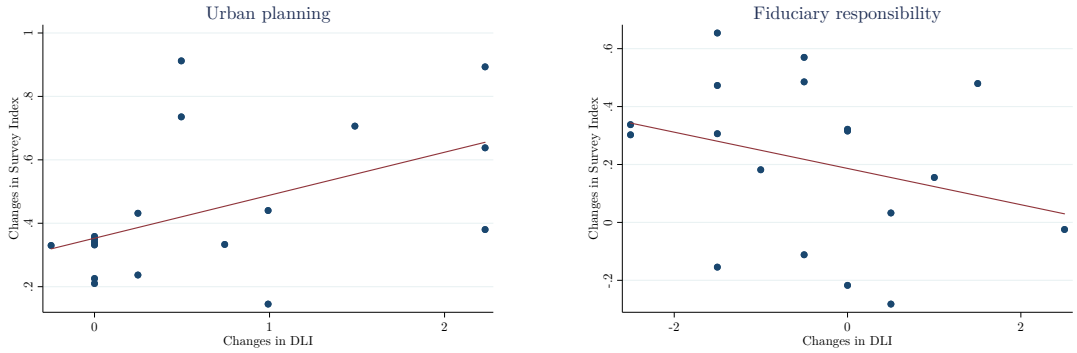
	Change APA	Change IE Survey
Accountability	0.23	0.04
Fiduciary	-0.42	0.21
Improved infrastructure	-0.57	0.41
Urban planning	0.67	0.44

Note: This table reports changes in APA scores between the 4th and 6th APAs (column (1)) and survey indices between first and second rounds (column (2)), for APA component and similar survey index. Changes are normalized to be interpreted as standard deviations of baseline. Indicator for "Increased revenue" is not presented since it was not collected on the 6th APA.

We present how variations in APA scores and equivalent household survey measures vary between the 2016-2018 period in the 18 LGAs receiving ULGSP in Figure B3. For the infrastructure and urban planning dimensions, we observe a positive correlation - although with a fair amount of noise as should be expected from 18 observations. The linear relationship between the two variables is also positive for accountability, but the scatter plot makes clear that most LGAs do not observe any variation in the DLI scores. Finally, for the fiduciary responsibility dimension we actually observe a negative correlation between changes in DLI and survey measures. Additionally, we present correlations for baseline, endline and changes in subcomponents for household survey measures in Figure B4.

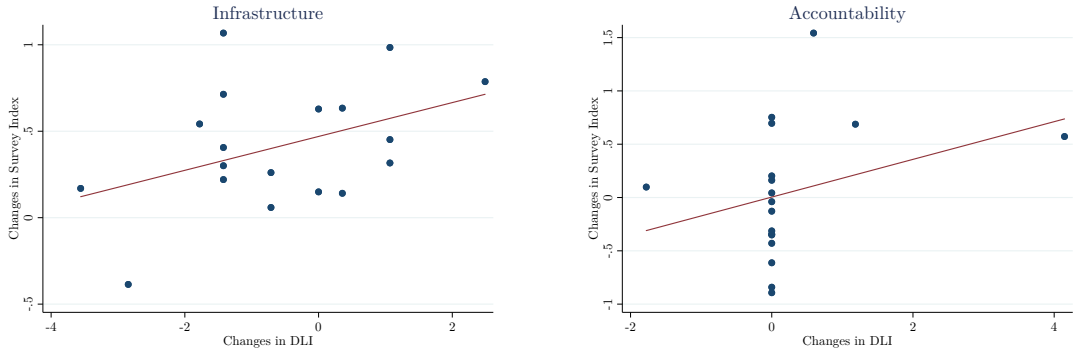
We also present correlations between APA scores and government officials' survey indices in Figure B5.

Figure B3: Changes in APA and Household Survey indices (2016-2018)



(a) Urban planning

(b) Fiduciary responsibility

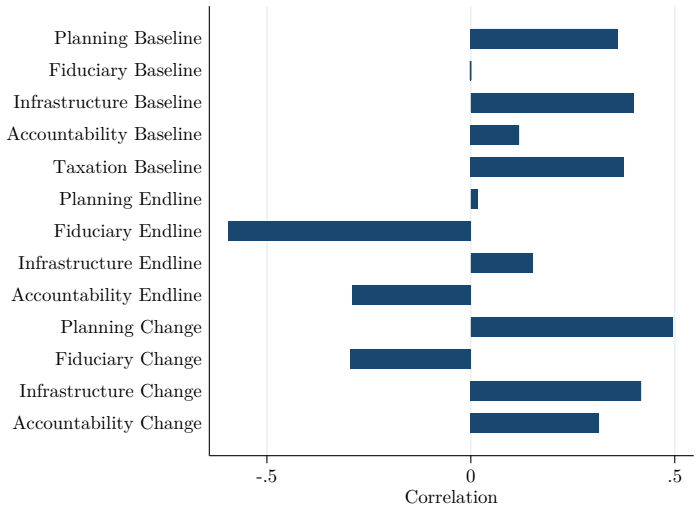


(c) Infrastructure

(d) Accountability

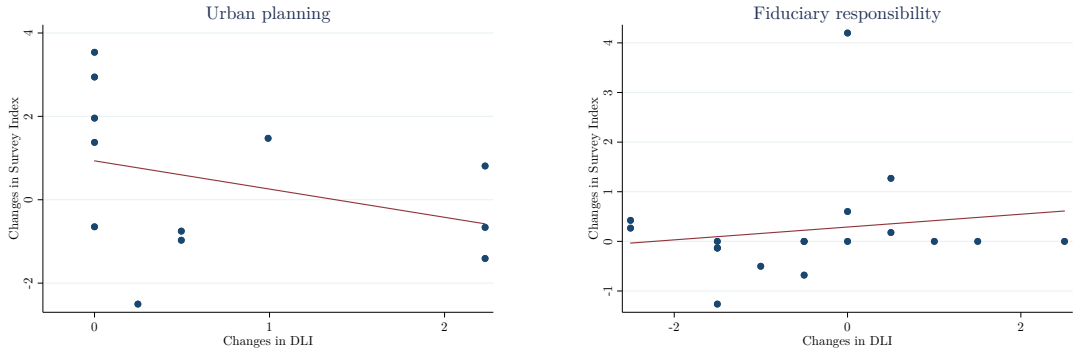
Note: This figure presents a scatter of changes in DLI subcomponent scores (x-axis) and survey indices (x-axis) by LGA. DLI scores are normalized so that changes can be interpreted as standard deviations in the distribution of the 4th APA scores.

Figure B4: Correlation between APA and household survey measures



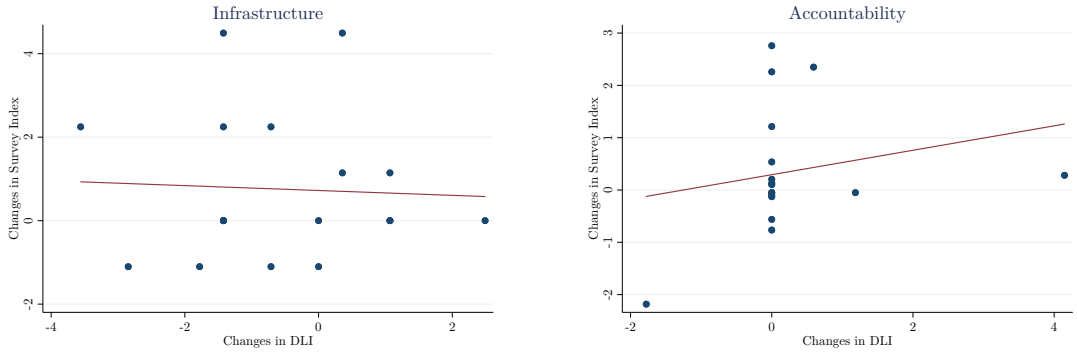
Note: This figure presents correlations across the 18 ULGSP LGAs for baseline, endline and changes in outcomes. Endline and changes are not reported for the taxation subcomponent since it was not collected in the 6th APA.

Figure B5: Changes in APA and Government officials' Survey indices (2016-2018)



(a) Urban planning

(b) Fiduciary responsibility

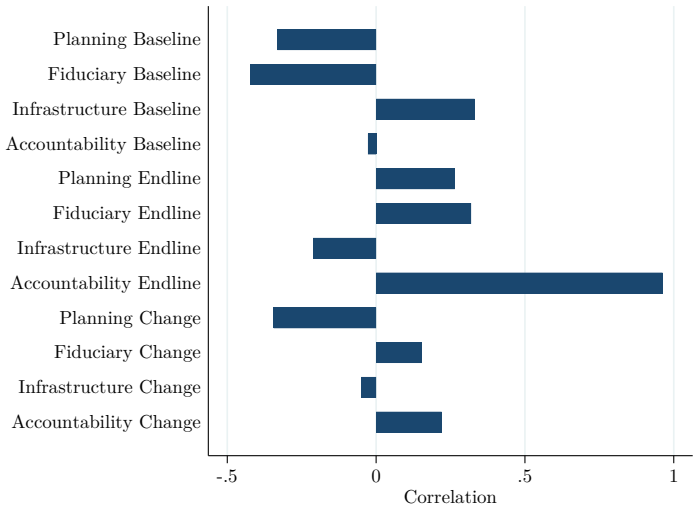


(c) Infrastructure

(d) Accountability

Note: This figure presents a scatter of changes in DLI subcomponent scores (x-axis) and survey indices (x-axis) by LGA. DLI scores are normalized so that changes can be interpreted as standard deviations in the distribution of the 4th APA scores.

Figure B6: Correlation between APA and government officials' survey measures



Note: This figure presents correlations across the 18 ULGSP LGAs for baseline, endline and changes in outcomes. Endline and changes are not reported for the taxation subcomponent since it was not collected in the 6th APA.

C Pre-trend checks

The key identifying assumption of our difference-in-differences strategy is that, in the absence of the ULGSP, outcomes in treated and reference units would have evolved similarly. While this parallel trends assumption is not directly testable, in this section we use nighttime light (NTL) data as proxy for economic activity and assess whether treatment and reference units presented similar trends before our first round of interviews.

In order to cover a longer time period before the intervention, we use the recently harmonized NTL dataset developed by Li et al. (2020), covering the 1992-2018 period. This is a synthetic dataset that uses the original Defense Meteorological Satellite Program (DMSP) data for 1992-2013 and calibrate data from the Visible Infrared Imaging Radiometer Suite (VIIRS) for the period 2012-2018. Following recommendations from the authors, in our main specification we drop any pixels with NTL value lower than 7.[30]

We present results in Figure C1. In panel A we present average NTL, with 95% CI, for ULGSP and comparison wards included in our survey sample. Consistent with the fact that ULGSP was targeted at more urban districts, the average nighlights are higher for those wards when compared to comparison. The trends before the first assessment, however, do not suggest differential pre-trends. We test this formally using a difference-in-differences model of the following form:

$$ntl_{wy} = \alpha + \gamma_w + \theta_y + \sum_{y=1999}^{2018} \beta_y (ULGSP * year)_{wy} + X_{wy} + \epsilon_{wy}$$

where ntl_{wy} are measures of nighlight in ward w in year y, γ_w and θ_y are ward and year fixed-effects, and the coefficients β_y measure the differential nighlights in ULGSP wards in each period. We allow for differential linear trends by region, included in the time-varying vector X_{wy}. Since treatment is defined at the LGA(district)-level, we cluster standard errors at that level. We use the entire sample period 1992-2018, but pool all years before 2000 in one coefficient.

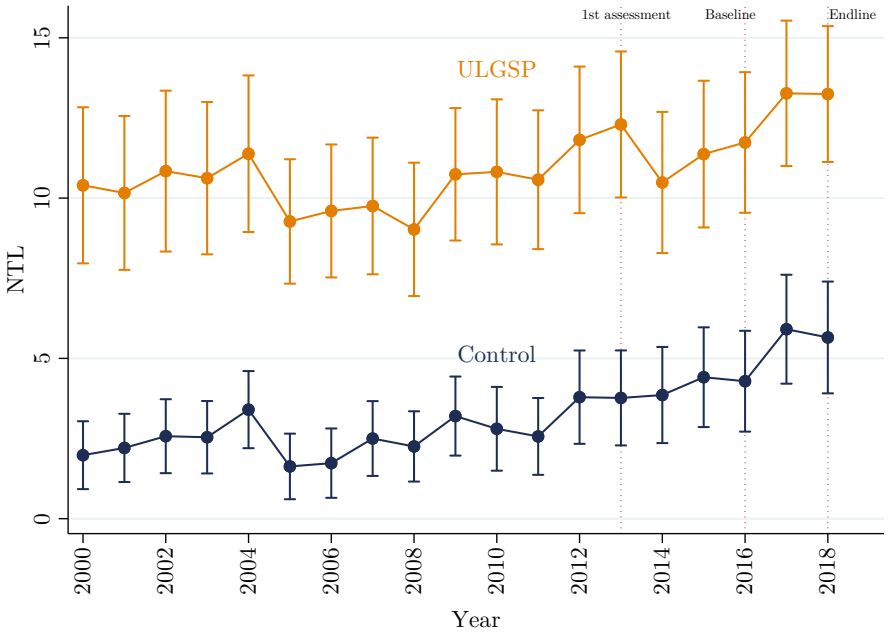
We present the β_y coefficients in panel B of Figure C1. While results are somewhat noisy, they do not suggest a differential trend between ULGSP and comparison before the 1st ULGSP assessment in 2013. While there is a temporary decrease in nighlights in ULGSP wards relative to comparison ones in 2014, that differential is temporary and quickly disappears in the following years.

We also present results using not only wards included in our survey, but measuring average nighlight at the entire district in Figure C3. Here we observe some degree of convergence between 2000 and 2013: the difference in mean nighlights in comparison and treatment districts falls by about 2 NTL points in the period (in 2011 the standard deviation (s.d.) of NTL across the approximately 800 wards included in study districts was 6.2, meaning that the estimated convergence in NTL was smaller than 0.3 s.d.). This is reflected in the DiD coefficients in Panel B. A

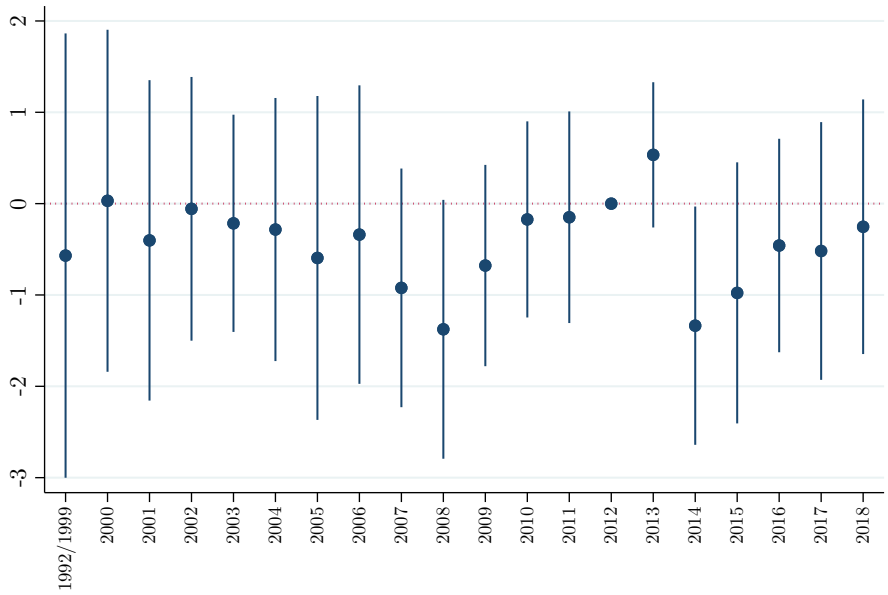
[30] In Figure C2 we present the original series and the adjusted series, as well as raw VIIRS data for the period 2012-2018.

similar pattern is observed in Figure C4, where we plot levels of nighlight and DiD estimates for the ward with strongest nighlights in 2011, as a proxy for the urban center in each district. We also observe some degree of convergence, particularly in the early 2000s.

Figure C1: DID estimates



(a) Means and 95% CI (2000-2018)

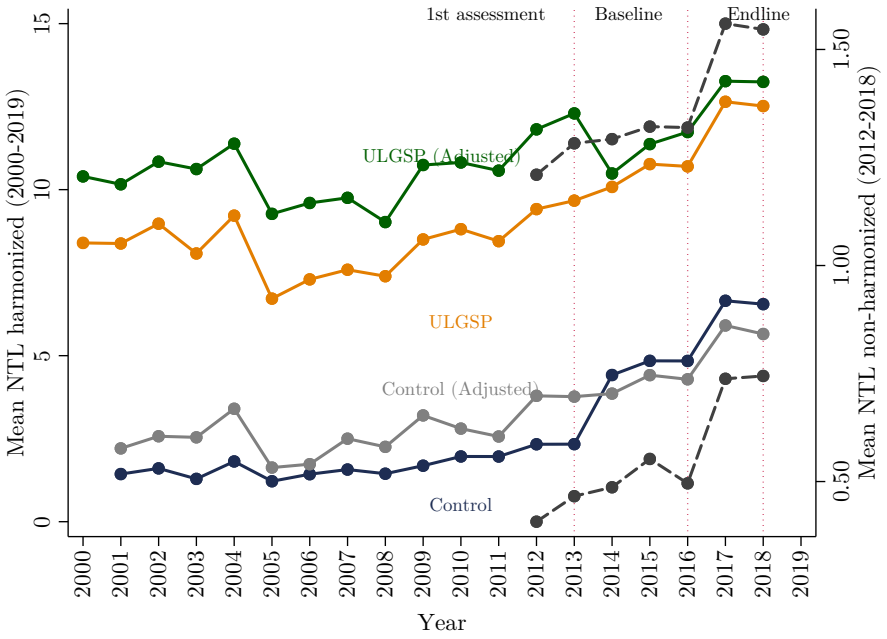


Note: Sample size is N = 3054

(b) difference-in-differences estimates

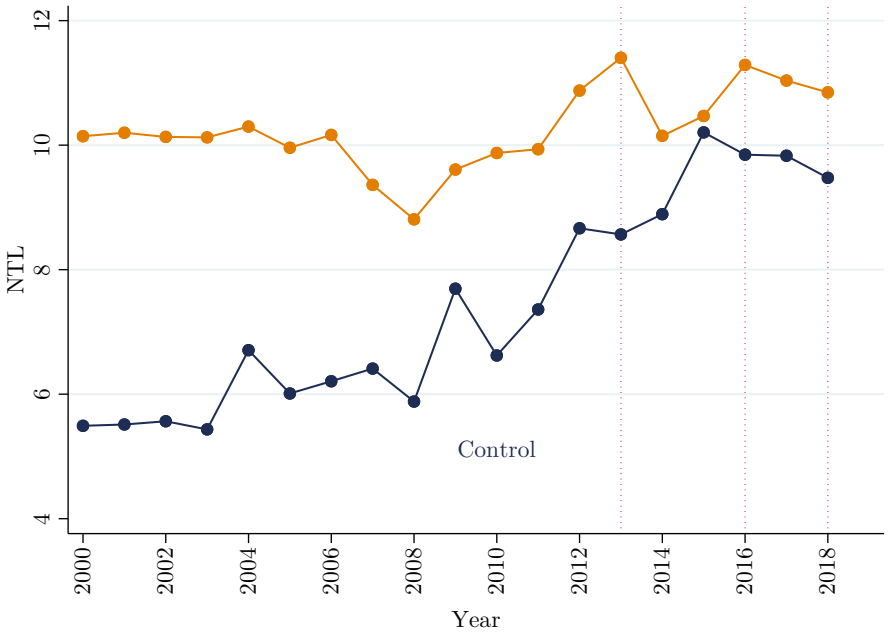
Note: Panel A presents mean NTL (with 95% CI) for wards in ULGSP treatment and comparison districts. The sample is restricted to wards included in the household and government officials surveys. Nighlight data excludes pixels with NTL < 7. Panel B presents results from DiD estimates. Coefficients obtained estimating equation $ntl_{wy} = \alpha + \gamma_w + \theta_y + \sum_{y=1999}^{2018} \beta_y (ULGSP * year)_{wy} + X_{wy} + \epsilon_{wy}$. Standard errors are clustered at the LGA, the level of treatment.

Figure C2: Average nighlights in treatment vs. comparison wards

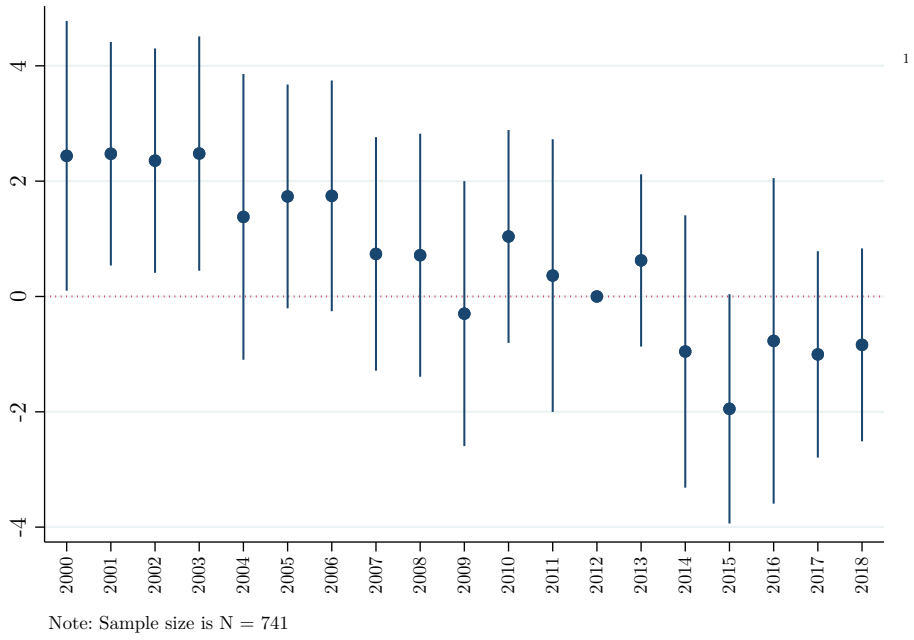


Note: This figure presents average NTL for ULGSP and comparison wards using different NTL measures. Adjusted series use harmonized data excluding pixels with NTL < 7, following the recommendation in Li, Zhou, Zhao and Zhao (2020), while the unadjusted series use all pixels. The dashed lines show data from the raw VIIRS dataset, covering only the 2012-2018 period. The sample is restricted to wards included in the household and government officials surveys.

Figure C3: DID estimates - NTL measured at district level



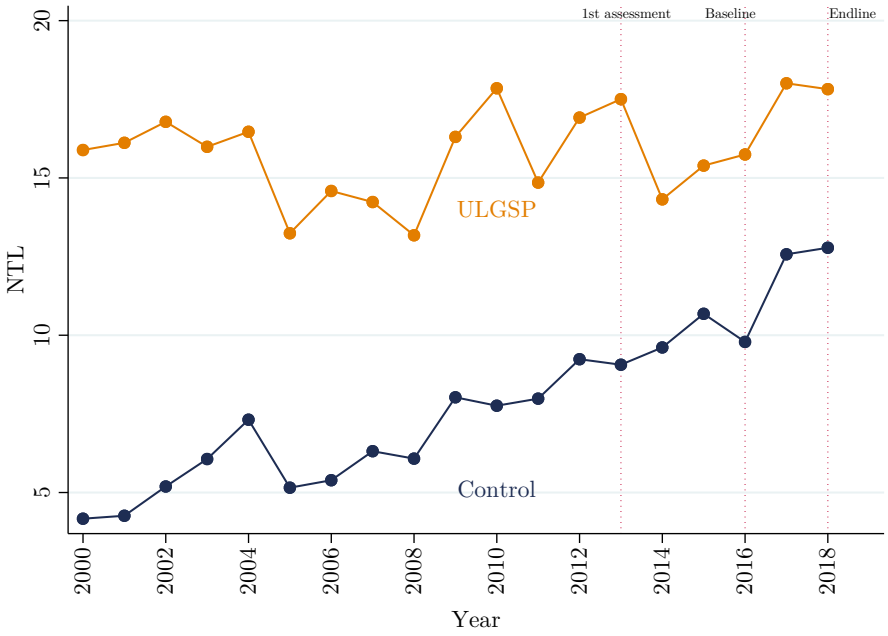
(a) Means (2000-2018)



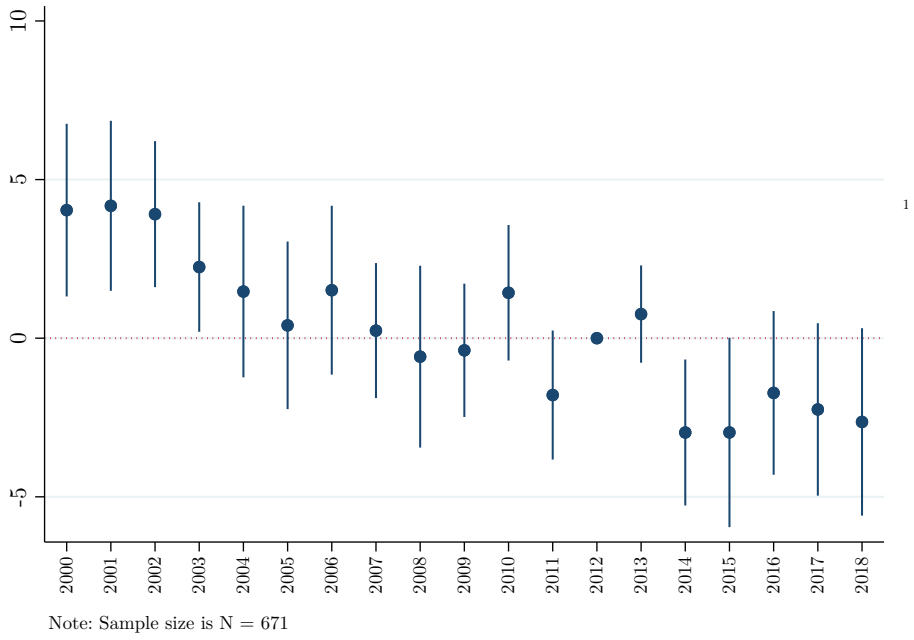
(b) difference-in-differences estimates

Note: Panel A presents mean NTL for districts in ULGSP treatment and comparison districts. Nighlights are measured as average in each of the study districts. Nighlight data excludes pixels with NTL < 7. Panel B presents results from DiD estimates. Coefficients obtained estimating equation $ntl_{wy} = \alpha + \gamma_w + \theta_y + \sum_{y=1999}^{2018} \beta_y (ULGSP * year)_{wy} + X_{wy} + \epsilon_{wy}$. Standard errors are clustered at the LGA, the level of treatment.

Figure C4: DID estimates - for ward with strongest nighlight in 2010 in each district



(a) Means and 95% CI (2000-2018)



(b) difference-in-differences estimates

Note: Panel A presents mean NTLfor wards in ULGSP treatment and comparison districts. The sample includes only the ward with strongest nighlight in 2010 in each of the study districts. Nighlight data excludes pixels with NTL < 7. Panel B presents results from DiD estimates. Coefficients obtained estimating equation $ntl_{wy} = \alpha + \gamma_w + \theta_y + \sum_{y=1999}^{2018} \beta_y (ULGSP * year)_{wy} + X_{wy} + \epsilon_{wy}$. Standard errors are clustered at the LGA, the level of treatment.

D Appendix Tables

Table D1: Government Officials - Diff-in-Diff alternative specifications

	Master Plan Up to date	Internal Audit independent	Capacity to raise local revenue	Staff capacity Index	Management Index	Performance Culture Index
			Panel A: Respondent fixed-effects			
DiD Coefficient	0.056	-0.281	-0.525	0.117	-0.127	-0.003
	(0.165)	(0.394)	(0.313)	(0.165)	(0.205)	(0.274)
Observations	240	888	932	572	810	894
			Panel B: Semiparametric Difference-in-Difference Estimator			
DiD Coefficient	-0.011	-0.282	-0.550***	0.039	-0.129	0.006
	(0.131)	(0.189)	(0.141)	(0.129)	(0.135)	(0.133)
Observations	119	443	465	285	404	446

Note: This table reports regressions using each of the described indices as dependent variable. Panel A reports results from a specification including respondent fixed-effects, while Panel B reports estimates using Abadie's Semiparametric DiD Estimator (Abadie, 2005). The smaller number of observations compared to the tables in the main text is due to the fact that once we use fixed effects we drop all observations for which either baseline or endline indices are missing. Regressions in panel B use the change in outcome as dependent variable, so the number of observations is half that of the panel data. Selection into treatment is balanced for respondents' age, gender and wealth index. Standard errors clustered at the LGA level are reported in parentheses (* p<0.1, ** p<0.05, *** p <0.01)

Table D2: Households - Diff-in-Diff alternative specifications

	Urban Planning Systems	Fiduciary Responsability	Infrastructure Management	Accountability & Transparencey	Views on Taxation and Fees
			Panel A: Fixed Effects		
DiD Coefficient	0.070	0.117	0.091	0.057	0.058
	(0.085)	(0.093)	(0.131)	(0.251)	(0.103)
Observations	4984	2390	1512	274	3416
R-Squared	0.579	0.607	0.648	0.555	0.590
			Panel B: Semiparametric Difference-in-Difference Estimator		
DiD Coefficient	0.091*	0.134*	0.093	0.071	0.039
	(0.053)	(0.074)	(0.085)	(0.240)	(0.062)
Observations	2490	1195	754	137	1706

Note: This table reports regressions using each of the described indices as dependent variable. Panel A reports results from a specification including respondent fixed-effects, while Panel B reports estimates using Abadie's Semiparametric DiD Estimator (Abadie, 2005). The smaller number of observations compared to the tables in the main text is due to the fact that once we use fixed effects we drop all observations for which either baseline or endline indices are missing. Regressions in panel B use the change in outcome as dependent variable, so the number of observations is half that of the panel data. Selection into treatment is balanced for respondents' age, gender and wealth index. Standard errors clustered at the LGA level are reported in parentheses (* p<0.1, ** p<0.05, *** p <0.01)