

Influence in Rural India: An Experimental Approach

Simon Chauchard

Simon.Chauchard@dartmouth.edu

Neelanjan Sircar

nsircar@sas.penn.edu

August 1, 2016

1 Questions

Indian voters have long been thought to make collective decisions ([Chandra, 2004](#)), especially in rural areas. This suggests that they are influenced not only by their immediate kinship networks ([Sircar, 2015](#)), but also by local actors beyond their kinship network. In rural India, a whole array of local characters plays such a role during electoral campaigns ([Björkman, 2014](#); [Chauchard, 2016](#)), as candidates and party networks rely on a variety of locally known, influential citizens to draw crowds to meetings, canvass and eventually attempt to convince voters. These individuals also play a role between elections, as they serve as intermediaries between elected officials and citizens, assisting citizens access the state and assisting elected officials with the mobilization of citizens and with the implementation of their policies.

While they have sometimes been clubbed under one label, as "brokers" or "intermediaries", these local-level actors are astonishingly diverse. The literatures on distributive politics, "claim-making" and electoral campaigns in India ([Manor, 2000](#); [Krishna, 2002](#); [Auerbach, 2013](#); [Dunning and Nilekani, 2013](#); [Bussell, 2014](#); [Björkman, 2014](#); [Kruks-Wisner, 2015](#); [Chauchard, 2016](#)) suggest that these actors may be elected or unelected, partisan or non-partisan, that they may belong to a variety of caste and ethnic groups, and that they may be more or less proximate to voters on a number of different dimensions.

As a result of this diversity, we still know very little about the identity, the role and the relative ability of these crucial actors. In light of this lack of data, and in light of the diversity of these individuals, this project has several objectives, which we will likely explore in several different papers (or more likely, in several different chapters of a book manuscript):

1. Generate descriptive data on the profile of these local actors.
2. Document the role that these different types of locally influential actors play in village life. Specifically:
 - (a) Determine which influencers better know villagers. Relatedly, determine which villagers does each type of local influencers know best and which villagers does each type of influencer have information on.
 - (b) Make inferences as to which villagers each of these local actors are willing to help and assist in their interactions with the state.
3. Finally and most importantly, determine which of these influencers are better at mobilizing and persuading citizens. That is, determine which individual-level characteristics of these local "influencers" makes them the most effective at mobilizing and persuading citizens, and in reverse, from which kinds of influencers voters are most likely to take their cues at the local level (This also allows us to determine the kinds of voters that different subtypes of influencers are likely to be able to mobilize and/or persuade).

This PAP details our research strategy to make inferences on all of these questions.

The "hypotheses" section below outlines specific hypotheses relevant to the more causal and quantitative parts of the project, i.e. points 2a, 2b and 3 above.

2 Motivation and Contribution

Inferences on the way in which local actors influence citizen in rural India are relevant to at least three different literatures. It should first improve the relatively thin literature on preference formation during electoral campaigns in India ([Chandra, 2004](#); [Sircar, 2015](#)). While this literature has frequently acknowledged the role of collective decision-making and of group-level coordination in preference formation ([Srinivas, 1955](#)), we still know very little about the way in which these mechanisms concretely operate on the ground, or about the types of individuals who are best able to coordinate divergent interests on the ground in India.¹ Our inferences should thus be of interest to both scholars of political campaigns and to practitioners in charge of organizing these campaigns and in

¹This is also true of the comparative literature on this question. Economists and marketing experts have long researched the drivers and the effects of persuasive communication (see review in [DellaVigna and Gentzkow \(2010\)](#)). Closer to the focus of this study, political scientists have researched the effects and the drivers of political communication ([Lazarsfeld 1944](#), [Gerber et al 2007](#)), the effectiveness of political advertising ([Gerber et al., 2007](#)), the effects of get-out-the-vote efforts ([Green and Gerber, 2008](#)), and the effects of the news media on voters ([DellaVigna and Kaplan, 2007](#); [Gentzkow, 2006](#)). Yet, while a handful of studies touch on this question ([Humphreys, Masters and Sandbu, 2007](#); [Dewan, Humphreys and Rubenson, 2014](#)), political scientists have so far not thoroughly explored the impact of individuals and the impact of the individual-level characteristics of local intermediaries on mobilization and persuasion.

charge of recruiting competent ground-level troops for their candidates, echoing recent studies ran in other contexts (Enos and Hersh, 2015; Dewan, Humphreys and Rubenson, 2014). Furthermore, the inferences made in this project should allow us to determine which kinds of influencers should be recruited in order to mobilize and persuade specific subgroups of voters.²

This research is also relevant to the literature on politician-voters relations in emerging democracies. The comparative literature on clientelistic networks (Stokes et al. 2014) has long highlighted the role of partisan brokers. These intermediaries are known to perform a variety of brokerage and mobilization related tasks, in a variety of contexts (Magaloni, 2006; Stokes et al., 2013; Hicken, 2009; Kitschelt and Wilkinson, 2007; Van de Walle, 2007). But in India (Chauchard, 2016; Berenschot, 2014; Björkman, 2014), as elsewhere (Holland and Palmer-Rubin, 2015; Larreguy, Montiel and Querubin, 2016), these partisan actors coexist with many other types of intermediaries or informal actors. While many of these brokers have partisan affiliations, many others derive their power from other traits such as shared ethnicity with citizens or other existing social relations. Indeed, much of the early work on patronage in Indian democracy, under the so-called "Congress System" (Srinivas, 1955; Weiner, 1967), focused on how the governing political party had to co-opt traditional local elites in order to stay in power. The implication is that parties may be beholden to these local elites, not the other way around, as is often assumed in the clientelism literature. Besides, further examination in the Indian context has found a significant democratization of the space of brokers, away from the traditional local elites that characterized early Indian democracy (Manor, Krishna). Instead of simply relying on long-term party agents, elites thus often engage in more opportunistic alliances with leaders of non-partisan organizations or with influential citizens who are only loosely affiliated to candidates in the short-term, around elections. Citizens, on the other hand, often require the assistance of local intermediaries who are not always party workers, as evidenced by the growing literature on claim-making (Manor, 2000; Krishna, 2002; Auerbach, 2013; Kruks-Wisner, 2015).³

Our research builds on these literatures in a number of ways. In order to determine which individual-level characteristics of influencers matter, we compare the effectiveness of the various actors that appear in this literature. Besides, as we do this, we determine the extent to which various intermediaries are able to influence citizens whom they assist, and hence the extent to which the relationship between these intermediaries and citizens is truly a quid pro quo relationship.

Finally, this research is relevant beyond politics. A growing literature is interested in the way in which useful information and virtuous behaviors can spread thanks to particularly influential individuals. In this spirit, scholars interested in the diffusion of health-related behaviors, micro-finance and other issues have investigated the extent to

²While long-term party intermediaries may for instance serve to mobilize core supporters of the party, non-partisan intermediaries can mobilize and convince a more diverse set of voters.

³ These informal actors have been described as fixers, dalals, or naya netas, among other terms.

which social networks can be used to make policy interventions more effective (Banerjee et al., 2013; David A Kim et al., 2015). These interventions have now shown that targeting network-central individuals can lead to virtuous behavioral cascades. While our approach is less ambitious, we merely identify influencers by asking a subset of our target population whom they might be rather than by mapping entire networks, we build on this research program by identifying the characteristics of individuals that are likely to be the most influential, and hence the most likely to trigger behavioral change.

From a theoretical standpoint (detail on this can be in the section on hypotheses below), a wealth of characteristics of "influencers" may make it more likely that they successfully mobilize a given individual. Among other characteristics, the following ones should for instance matter:

- **Ethnicity:** whether the influencer is from a dominant group and/or whether the influencer is from the same group than the individual she attempts to mobilize.
- **Partisan Background:** whether the influencer is from the ruling party at the state level and whether the influencer is from locally dominant party should matter. Whether the influencer is from the individual's favorite party should also matter
- **Electoral Status:** whether the influencer is or was elected at the local level likely matters
- **Class:** whether the influencer is wealthy/educated, whether the citizen that the influencer attempts to mobilize is wealthy/educated, and whether the influencer belongs to the same class than the individual should all matter.
- **Reciprocity:** Whether the influencer has in the past helped the individual he attempts to influence (or whether the target of influence has relied or relies on the influencer to access benefits or income) should matter.
- **Strength of ties:** how closely connected the influencer and the individual are should also matter.

Our design – detailed below – allows us to make inferences on the extent to which each of these characteristics matter and on the subpopulations of citizens that influencers with each of these diverse characteristics are likely to successfully mobilize.

3 Research Design

This project focuses on understanding how organizations and candidates can mobilize citizens at *the most local* level, i.e., the level at which organizations, parties or candidates attempt to have direct face-to-face contact with voters. In this environment, local intermediaries (hereafter referred to as "influencers") and voters tend to have a lot of information about each other because they are neighbors or because their personal networks

overlap.

Data collection takes place over three days in each polling booth we target and takes place in several steps. The following subsections properly detail each of these steps chronologically (DAY 1, DAY 2, DAY 3). Each instrument (either a word document or an excel spreadsheet) mentioned in this document is denoted in violet-colored font, **as such**.

Before we detail this design, the next subsection details the sampling strategy for selecting polling booths.

3.1 Sampling of Polling Booths

The project combines surveys and experiments in 176 rural polling booths located across three districts of the Indian state of Bihar in order to measure the influence of different types of influencers. In this subsection, we detail our sampling strategy.

We first purposely selected 3 mostly rural districts of Bihar: Buxar, Nalanda, and Vaishali.⁴ These were chosen to ensure some minimal levels of cultural, political and socio-economic diversity in our sample. Nalanda is located about 50-100 km south of Patna (the state capital) and is a Magahi-speaking area. Buxar is located about 125-200 km west of Patna and is located in the Bhojpuri-speaking area of the state. Finally, Vaishali is located just across the Ganges River, north of Patna, and is located in the Maithili-speaking area of the state.

In each of these three districts, three blocks were subsequently selected. In selecting these blocks, we mostly had security and feasibility concerns in mind. In each district, we first made sure to exclude blocks that would be prone to flooding (which would have made the work of the research team complicated during the rainy season). Second, more generally speaking, we excluded several blocks that were not easily accessible by road. This was to ensure the security of survey teams as well as to guarantee that our implementing partner (SUNAI) would be able to implement the complex protocol detailed below in a timely fashion. Among remaining blocks in each district, we then randomly selected three blocks.

In each of these blocks, we randomly selected booths. To do this, we selected the booths using a variant systematic sampling.⁵

⁴The state of Bihar is almost 90% rural. The traditional Indian village is split into neighborhoods, or *tolas*, that are typically composed of a single caste group (*jati*) or a non-Hindu religious group. The lowest defined unit in the political structure of India is the polling booth (PBA), which is typically composed of several *tolas* (Note that there may be several polling booths in a single village).

⁵That is, we broke the list of polling booths in each block into 40 intervals with approximately the same number of polling booths, randomly selected whether we would take odd or even numbered intervals (i.e., first, third, ..., or second, fourth, ...), and then we randomly selected a polling booth in each interval. Each interval contained consecutive "polling booth numbers" which means that they are likely spatially

3.2 Day 1

In each targeted PBA, the research team starts by identifying 3 individuals who will serve as influencers in the rest of the experiment.

The identification of the first two influencers takes place on the morning of Day 1. The first two selected influencers (T1 and T2) are those whom a random sample of voters identifies as most influential in the village, as described in section 3.2.1; these are individuals who are plausibly skilled at influencing voters. The third individual is a randomly drawn male voter in the village, as described in section 3.3.2. The randomly drawn male voter (T3) provides a benchmark against which we can assess the skill of the two influencers selected by voters (T1 and T2).

3.2.1 Identifying "Real Influencers"

The research teams first informally speaks to villagers in gathering spots across the village to determine a list of possible influential individuals.

The objective is to get a diverse set of names of people whom others think might be influential. The three questions the research team systematically asks to each individual (or groups of individuals) they meet are:

1. *Who among residents of this area is most influential?*
2. *When it comes to social issues, whose opinions do people listen to the most around here?*
3. *When people seek to solve small problems outside the family in this village without approaching the panchayat or political party, who do they go to?*

The research team asks these three questions in at least five different locations within the PBA. These areas should be relatively dispersed within the PBA.

The research team obtains between 10 and 15 different names at the end of this process. If they have fewer than 10 names, they keep on visiting more locations until they have 10 names.

As they obtain names in response to these questions, the research team asks for a few additional details about each of the individuals named (phone number, position in party, elected position, profession, age, community/social group).

The research team enters their responses in a document based on the attached excel spreadsheet titled **LIST OF INFLUENCERS (DAY 1)**.

The research teams then visit a random selection of 12 households within the polling booth. The objective of this interview is to ask a sample of citizens to identify which of the names (from list we have collected) they believe to be most influential, in addition to clustered.

background information, which will be necessary to the continuation of the project on day 2.

To identify these voters, the research team relies on a sampling document titled **CITIZENS SAMPLING_PBAnumber (DAY 1)**. Importantly, these citizens are drawn from a different sample than the citizens we interview on day 2 and day 3.⁶ The spreadsheet contains a list of voters, as well as 19 replacements. Once they have reached a voter's household, surveyors attempt to interview the voter mentioned on **CITIZENS SAMPLING_PBAnumber (DAY 1)**. In case that voter is female, they ask to speak to a male member of the household to establish who should be interviewed (in case the voter selected on the spreadsheet is not male, the research team attempts to interview a male member of the household, and not the respondent). In order to do so, the team makes a list of male members of the household, and uses a Kish table strategy to randomly select a male member.⁷

The brief interview that then takes place is described in **CITIZENS' OPINIONS on INFLUENCERS (DAY 1)**. As mentioned above, this interview allows us to obtain fine-grained background information on the respondent as well as an estimate of the relative power and influence of the various influencers whose names are listed on **LIST OF INFLUENCERS (DAY 1)**.

Questions 19.1 and 19.2 are used in **CITIZENS' OPINIONS on INFLUENCERS (DAY 1)** to assess the popularity of each influencer in our collected list, and select for the next day.

3.3 Day 2

3.3.1 Selecting Two "Real Influencers" From the List

To select two influencers (T1 and T2) from the list we follow the following steps on the evening of day 1.

1. Teams supervisors create a randomized list including all the names.
 - (a) To do this, the supervisor writes each name on a different slip of paper. He then puts the slips in a bag and draws them one by one. The first name goes

⁶In order to minimize overlap, we break the voter list in a polling booth into approximately 48 evenly sized intervals of contiguously numbered voters. Every odd or even interval is selected again. From these 24 intervals, again either odd or even intervals are selected. If the odd intervals are selected for the first day, then the remaining intervals are used for the rest of the sample (and vice versa). This guarantees that there is little to no overlap in sampled households across days (voters are replaced if the same household is sampled across days). Once intervals are selected for a day, a voter is selected at random from within each interval.

⁷Given the short timeline, in practice this will be a male household member who is home on this day.

on row 1 of **Randomized list of Influencers**, the second on row 2, the third on row 4, etc ...

- (b) For each name, the supervisor copies the tally of votes (the number of times each name was chosen as a response to question 19.1 or 19.2 in **CITIZENS' OPINIONS on INFLUENCERS (DAY 1)**). They also code whether each individual named on the list is or was just elected in local institutions (They enter Y if currently elected or elected during the 2011-16 period, N if not currently elected nor elected during last term).

2. Selection of T1.

- (a) To select our first influencer, we simply select the most popular individual on the list. We then attempt to get his consent and to interview her/him using **INFLUENCER QUESTIONNAIRE (DAY 2) V1**. We go down the list in order of popularity if he is unavailable or does not provide consent. If there is a tie in popularity, the research team picks the individual that was randomly placed higher on the list among the most (equally) popular. Ex: two influencers have 6 votes. Supervisor picks the one whose name is in the highest row on the **Randomized list of Influencers**.

3. Selection of T2.

- (a) To select a second influencer, the supervisor modifies **Randomized list of Influencers** and creates **Modified-Randomized list of Influencers**. To do this, he crosses out every name of people that were already selected as T1, or names of people we could not manage to find when looking for T1. He also crosses out every name of people that are currently elected or were elected during 2011-2016. This allows us to ensure that we do not only select influencers who currently are (or just were) elected. He then copies the remaining names in the same order on **Modified-Randomized list of Influencers**.
- (b) He now selects the most popular individual (i.e., individual with most votes) in **Modified-Randomized list of Influencers**. If there is a tie as to which individual is most popular, he picks the individual higher on the list among the most popular.

Overall we thus always pick two different individuals, and at least one individual who is not elected. The randomization scheme described above guarantees that those potential influencers receiving the same number of votes from day 1 citizens have an equal probability of being selected.

3.3.2 Identifying the 3rd Influencer

The 3rd influencer in each PBA is the head of household of a randomly drawn household within the PBA. The randomly drawn male head of household provides a benchmark against which we can assess the skill of the influencers selected by voters. In each case, we provide the research team with a random list of voters' names (and replacements) to draw from (the relevant sampling document is titled **SAMPLING 3rd INFLUENCER_PBAnumber DAY 2**). In case the voters selected on the spreadsheet are not heads of households, the research team should attempt to interview the head of household, and not the respondent.⁸

3.3.3 Securing Influencers' cooperation and Influencer Interview

Once the research team has identified influencers (and potential replacements), the research team secures the cooperation of the 3 selected influencers (likely: morning of day 2) and asks them a few questions.

INFLUENCER QUESTIONNAIRE (DAY 2) describes how the research team proceeds upon contacting influencers.

As shown by the instrument itself, these influencers are asked to help the research team mobilize local citizens for a focus group about "what villagers need and what problems they are facing in this PBA" and to consent to a long interview (which includes a number of games).

Practically, we ask these individual to consent that we disclose to other citizens residing in the polling booth area that they recommend attendance to the event we organize on day 3. During this interview of selected influencers, we also ask them to endorse a number of other behavioral recommendations that we subsequently use as "treatments" in the experimental section of the second citizen survey. Namely, we ask influencers: 1) to recommend that citizens behave generously in dictator and public goods games; and 2) to recommend that citizens we invite bring some of their neighbors to the focus group meeting. Finally, we ask these influencers to make attitudinal recommendations over low-prior issues on which we also subsequently interview citizens.

If they agree to this, a survey incorporating a number of games takes place. The research team conducts surveys and executes games with each of the three listed influencers. The instrument for this is **INFLUENCER QUESTIONNAIRE (DAY 2)** (used along with **INFLUENCER ANSWER SHEET (DAY 2)**).

The goal of these exercises is to understand the demographic and social background

⁸A voter selected at random and 4 replacements are given. All voters within five contiguous numbers on voting list are removed from the rest of the sample (from which citizens for voter-level surveys are drawn). This minimizes the likelihood that someone is sampled from the household of this person.

of influencers, their underlying preferences, and how much they know about voters the research team interviewed on day 1. Three games allow us to measure 1. their willingness to assist each of the respondents interviewed on Day 1, 2. their self-reported ability to mobilize each of these respondents, and 3. their proximity to, or their level of knowledge about each of these respondents.

By mid-day on day 2, the research team will thus have secured the cooperation of two influencers from [LIST OF INFLUENCERS \(DAY 1\)](#) and of one head of household listed on [SAMPLING 3rd INFLUENCER_PBAnumber \(DAY 2\)](#). They will then interview them using [INFLUENCER INTERVIEW \(DAY 2\)](#).

3.3.4 Citizens' Interviews

Starting the afternoon of day 2 (that is, after the three influencers' interviews), we interview another 12 randomly selected males drawn from [CITIZENS SAMPLING_PBAnumber \(DAY 2\)](#), a different sampling document. We randomly interview 6 of these 12 randomly selected male heads of households the afternoon of day 2 and 6 more the morning of day 3. This is done to understand time effects in the mobilization exercise described below.

The sampling document [CITIZENS SAMPLING_PBAnumber \(DAY 2\)](#) associates each targeted respondent to a treatment group, ranging from 1 to 4.

Each group corresponds to which influencer we associate each respondent with, in the interview that follows:

- 1 corresponds to Influencer number 1 on the second tab of [LIST OF INFLUENCERS \(DAY 1 and 2\)](#).
- 2 corresponds to Influencer number 2 on the second tab of [LIST OF INFLUENCERS \(DAY 1 and 2\)](#).
- 3 corresponds to Influencer number 3 the second tab of [LIST OF INFLUENCERS \(DAY 1 and 2\)](#). (Benchmark influencer)
- 4 corresponds to NO INFLUENCER (control group)

The research team interviews these 12 randomly selected male heads of household, following the procedure described in [CITIZEN LONG INTERVIEW \(DAY 2 and 3\)](#).

Whenever our respondents have been randomly assigned to one of the influencer groups (groups 1-3 above), the research team mentions the recommendation of the influencer before they answer/play, as described in the instrument. In the case of our control group, we simply mention a similar recommendation (from 'a villager', who remains unnamed) before respondents answer.

See [CITIZEN LONG INTERVIEW \(DAY 2 and 3\)](#) for details on the content of the interview. Questions 26 to 29.1 provide us with some of the dependent variables we later use to measure influence. Questions 26 and 26.1 allow us to measure whether suggesting that an influencer recommends that the respondent come to the meeting (and bring neighbors with him) increases the likelihood that that respondent self-report an intention to do so during an interview. Question 27 (dictator game) allows us to measure whether an influencer' recommendation to be generous (i.e. to retain little of the original endowment) has an effect on the actual costly behavior of respondents. Question 28 gets at the same point, though in the context of a Public Goods Game. This in turn allows us to measure whether respondents believe that the behavioral recommendation of an influencer will have an influence on a majority of villagers interviewed. Finally, questions 29 and 29.1 allow us to measure whether suggesting that an influencer recommends participation in participatory development programs increases the likelihood that that respondent declares being in favor of these programs and being willing to sacrifice significant time to them.

Finally, the treatment for our main behavioral dependent variable is delivered during the interview, as each respondent is invited to attend a focus group meeting at the end of DAY 3. They receive an [INVITATION LETTER \(DAY 2 and 3\)](#) mentioning that fact (on which the respondent number is also included), as well as a total of 2 additional copies of [INVITATION LETTER \(DAY 2 and 3\)](#) that we ask them to distribute to their neighbors, as detailed in the instrument prior to question 26. As discussed above, we use the effect of this treatment on their actual attendance (which we measure in [INVITATION OBSERVATION \(DAY 3\)](#), as discussed below) to develop a behavioral and naturalistic measure of influence.

3.4 Day 3

3.4.1 Continuation of Citizen Interviews

In the morning, the research team completes the final 6 citizen long interviews (see [2.2 CITIZEN LONG INTERVIEW \(DAY 2 and 3\)](#)). These should be completed by 1pm on that day.

The individuals for survey are drawn from a separate sampling document titled [CITIZENS SAMPLING_PBAnumber \(DAY 3\)](#).

3.4.2 Remobilization efforts

About one hour before the meeting time (agreed upon in consultation with influencers on Day 2), the two enumerators revisit each of the citizens who were surveyed on day 2 and day 3 (that is, citizens from our second sample of citizens, all of which were

invited to the meeting). They will ask them once again whether they plan to come to the meeting, following the same prompts mentioned above. The protocol and the questions for this remobilization visit are described in a brief instrument called **REMOBILIZATION VISIT (DAY 3)**.

As seen in the instrument itself, enumerators also use this visit to measure whether citizens have passed on the letters of invitation, as suggested during the initial interview.

3.4.3 Focus Group Meeting

We end by conducting the focus group meeting announced on day 2 and day 3. The team leader leads the focus group according to the protocol described in **FOCUS GROUP PROTOCOL (DAY 3)**, while other enumerators take notes (see protocol).

FOCUS GROUP PROTOCOL (DAY 3) describes the protocol of the focus group, and the questions asked during the discussion. It also specified how we record which of the households visited on day 2 and 3 sent someone to the meeting. The research team does that by requesting to know who was given the **INVITATION LETTER (DAY 2 and 3)** of attendants at the beginning of the meeting (those who do not carry invitation letters can stay).

As mentioned on the protocol, we record invitations in a document titled **INVITATION OBSERVATION (DAY 3)**.

4 Hypotheses and Analyses for Paper 1: Which Influencers Better Know Villagers? (similar to point 2a above)

As mentioned above, the project described in section 3 has several objectives, which will be explored in several different papers (or alternatively, in several different chapters of a book manuscript):

1. Generate descriptive data on the profile of these local actors.
2. Document the role that these different types of locally influential actors play in village life. Specifically:
 - (a) Determine which influencers better know villagers. Relatedly, determine which villagers does each type of local influencers know best and which villagers does each type of influencer have information on.
 - (b) Make inferences as to which villagers each of these local actors are willing to help and assist in their interactions with the state.

3. Finally and most importantly, determine which of these influencers are better at mobilizing and persuading citizens. That is, determine which individual-level characteristics of these local "influencers" makes them the most effective at mobilizing and persuading citizens, and in reverse, from which kinds of influencers voters are most likely to take their cues at the local level (This also allows us to determine the kinds of voters that different subtypes of influencers are likely to be able to mobilize and/or persuade).

In the following sections (4 to 6), we outline hypotheses and specify analyses relevant to the more causal and quantitative parts of the project, i.e. points 2a, 2b and 3 above. In each case, we first detail how we measure our dependent variable(s). We then specify a series of hypotheses we will be testing as we attempt to answer each question. Finally, we detail modeling choices and analyses (specifying in each case each the statistical model we plan to rely on as well as the variables we plan to include in our analyses).

In this section (section 4), we start by describing hypotheses and analyses with respect to the knowledge that influencers have of the sampled voters (as in point 2a above).

For simplicity of exposition, we will standardize the predictive variables across these analyses. In particular, we will construct a matrix of variables, \mathbf{X} , consisting of the variables listed in rows xxx of the first tab (titled "2a variables") of [Variables and Measures declared in PAP](#). This document in each case refers to the concept measured by each variable and to the survey item (or items) on which we rely to construct each variable.

Influencer Characteristics

- a binary variable indicating whether the influencer was *not* a "benchmark" influencer (row 13)
- a binary variable indicating whether the influencer is currently elected (row 14)
- a binary variable indicating whether the influencer has been elected in the past (row 15)
- a binary variable indicating whether the influencer is from a locally dominant group or upper caste (row 16)
- a variable denoting the natural logarithm of the years an influencer has been affiliated with a specific political party (row 17)
- a binary variable denoting whether the influencer is affiliated with the locally dominant political party (row 18)
- a variable denoting the influencer's absolute deviation of an asset index from the mean value of the asset index among sampled voters (row 19 ⁹)

⁹In order to create the asset index, we fit a standard item response model on assets common across the influencer and voter survey. See below for an explicit model formulation of an item response model. The "ability" parameter of the model is used as the measure for the asset index. This is the same approach

- a variable denoting the influencer's absolute deviation of years of schooling from class 8 (row 20)
- a binary variable indicating whether the influencer is a member of non-political associations (row 21)
- a binary variable indicating whether the influencer has a high status job (row 22)
- a binary variable indicating whether the influencer has a bureaucratic or state-related job (row 23)
- a variable denoting the influencer's absolute deviation in age from the mean age of sampled voters in the polling (row 24)

Voter Characteristics

- a variable indicating the voter's absolute deviation in age from the mean age of sampled voters in the polling booth (row 26)
- a variable indicating the voter's absolute deviation in years of schooling from class 8 (row 27)
- a variable denoting the voter's absolute deviation of an asset index from the mean value of the asset index among sampled voters in the polling booth (the same asset index calculated above) (row 28)

Dyadic Characteristics

- a binary variable indicating whether the influencer is from the same social group as the sampled voter (row 30)
- a binary variable indicating whether the influencer is a co-partisan of the sampled voter (row 31)
- a binary variable indicating whether the influencer has helped citizen in the past (row 32)

The baseline for this analyses (when all predictors take the value of zero) will be a named influencer with the mean level of asset wealth among sampled villagers in the polling booth, mean age among sampled voters in the polling booth, and class 8 level of education who does not have political or associational membership, high occupational status, and who does not have partisan or caste/ethnic affiliation or domination over the voter. The baseline behavior will be measured towards a voter with class 8 level of education, the mean level of asset wealth among sampled voters in the polling booth, and mean age among sampled voters in the polling booth. All models are run in a Bayesian framework with diffuse priors.

used in Schneider and Sircar (n.d.).

In addition to these predictors, we will include a few polling booth level controls in our models in a matrix Z :

- the size of the village as per the 2011 Census of India
- the size of the polling booth as per the most recent voting list for the polling booth
- the number of *jati* (or religious minority) groups believed to be at least 10% of the population in the village
- distance to the district headquarters (in km)

As listed in rows 2-9 of the first tab of [Variables and Measures declared in PAP](#), our dependent variables of an influencer's knowledge of voters consist of the influencer's self-reported closeness to the voter (on an integer scale of 0 to 3), as well as a series of binary variables that characterize whether the influencer correctly guessed a number of attributes (caste, education, occupation, rooms in home, vote choice in 2015 state election, phone number) of the voter. Two of these binary variables, those that guess the educational level and the number of rooms in a voter's house, can be further characterized by "distance" from the correct value based on our data.

This yields three different types of models for the analysis. First, we run an ordered logistic regression model with the self-reported closeness to the voter on variables of interest. Second, we create an item response model over the binary guesses (correct/incorrect) of a voter's attributes, decomposing the parameter of an influencer's "ability" into a random effect and the impact of variable of interest. Finally, we run modified ordered logistic regressions using the distance from the correct value for guesses of education and number of rooms of the voter on variables of interest.

4.1 Hypotheses

As shown in rows 2-9 of the first tab of [Variables and Measures declared in PAP](#), we rely on several responses provided by influencers during the Game 3 section of [INFLUENCER QUESTIONNAIRE \(DAY 2\)](#) and on responses provided by citizens in [CITIZENS INTERVIEW DAY 1](#) in order to determine Influencers' levels of knowledge about citizens. We outline a series of hypotheses regarding influencers' level of knowledge about citizens. [Variables and Measures declared in PAP](#) specifically refers to the variable(s) used to test each of these hypotheses.

H1. "Real" influencers (T2 and T2) are more knowledgeable about citizens than "Benchmark Influencers" (T3).

Our second and third hypotheses relate to the elected status of the selected influencer:

H2. Influencers who are elected or were just elected (until the 2016 elections) in local village institutions know more about citizens than unelected ones.

- H3.** Influencers who were elected in the past in local village institutions know more about citizens than unelected ones.
- H4.** Influencers who are from a locally dominant Group or an upper caste know more about citizens than influencers from lower castes.
- H5.** Influencers know more about citizens from their own group than about citizens from other groups.

The next several hypotheses relate to partisanship and to the partisan background of influencers:

- H6.** On average, influencers who have been affiliated in the long run with a specific political party know less about citizens than influencers who have not been affiliated in the long run with a specific political party.
- H7.** Influencers who have been affiliated in the long run with a specific political party however know more about citizens who have also been affiliated to that party.
- H8.** Influencers who are affiliated with the locally dominant party (i.e. the party that won the seat in the 2015 state elections) better know citizens than influencers affiliated with other parties and than influencers affiliated to no specific party.
- H9.** Influencers who have the same preferences than citizens know more about these citizens than influencers affiliated with other parties and than influencers affiliated to no specific party.

The next hypotheses relate to the class and social status of the influencer and of the citizen. We measure class in two ways: as assets owned and as education.

- H10.** Influencers who are assets-wealthy know more about citizens than influencers who are assets-poor.
- H11.** Influencers who have a higher level of education know more about citizens than influencers with a lower level of education.

The next hypotheses relate to the occupation of influencers and to their membership in non-political organizations.

- H12.** influencers who are members of non-political organizations or associations know more about citizens than influencers who are not.
- H13.** influencers whose occupation is non-political but high status or with high-networking potential know more about citizens than influencers who are not.
- H14.** influencers who are bureaucrats, civil servants or have a state job know more about citizens than influencers who are not.

The final hypothesis in this section relates to the existence of a prior relationship between the influencer and citizen.

H15. influencers who have in the past helped a sampled citizen know more about that citizen than influencers who have not.

4.2 Models for Analysis

4.2.1 Model 1

Our first dependent variable of interest, *CLOSE*, is the influencer's self-reported closeness to the voter, which takes on non-negative integer, $q \in \{0, 1, 2, 3\}$. We use the notation \mathbf{X}_i and \mathbf{Z}_i to denote the i^{th} row of the respective matrices. We fit a multilevel ordered logistic model to data test our hypotheses of interest for voter $i \in I$, influencer $j \in J$ in polling booth $k \in K$ with random terms for the influencer, voter and polling booth.

$$P(\text{CLOSE}_{ijk} \leq q) = \text{logit}^{-1}(c_q - \mathbf{X}_i\boldsymbol{\beta}_X - \mathbf{Z}_i\boldsymbol{\beta}_Z - \alpha_k - \alpha_j - \alpha_i)$$

$$\alpha_i \sim N(0, \sigma_I^2); \quad \alpha_j \sim N(0, \sigma_J^2); \quad \alpha_k \sim N(0, \sigma_K^2)$$

$$q \in \{0, 1, 2\}; \quad c_0 \leq c_1 \leq c_2;$$

The parameter vectors $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_Z$ are used to test the hypotheses listed. Note that because probability of being less than or equal to the largest category is 1 by definition, we only need to model one less than the total number of categories. The parameters c_0, c_1, c_2 correspond to "cutpoints" between the categories of 0/1, 1/2, and 2/3, respectively. Finally, the α terms correspond to random effects for the voter, the influencer, and the polling booth in which the sample is being drawn. This model estimates the probability of selecting each category of the dependent variable, adjusting the probabilities to account for the predictors in \mathbf{X} . We note that a more general model is possible in which the cutpoints themselves vary as a function of i, j , and k . This is important, when, for instance, certain combinations produce higher percentages in the middle of plausible values for the dependent variable (i.e., 1 or 2). We will run such models as a robustness check.

4.2.2 Model 2

Our second dependent variable(s) are a series of six binary outcomes (correct/incorrect) asking for the influencer to report information about each voter. This sort of exercise has also been conducted in [Schneider \(2014\)](#) with respect to vote choice. In order to model the "ability" of each influencer to report information about voters, we fit an item response model that accounts for the ease of predicting each piece of information, as well as the ease of prediction for a particular voter and within a particular voting booth. For each piece of information $t \in T$ drawn from the space of pieces of information (T), the report from influencer $j \in J$ about voter $i \in I$ in polling booth $k \in K$ about information t

will be denoted as $t_{ijk} \in \{0,1\}$. We may then write down an appropriate item response model:

$$P(t_{ijk} = 1) = \text{logit}^{-1} (\alpha_{ik} + \alpha_j + \alpha_k + \mathbf{X}_i\beta_X + \mathbf{Z}_i\beta_Z - \theta_t)$$

$$\alpha_i \sim N(\mu_I, \sigma_I^2); \quad \alpha_j \sim N(\mu_J, \sigma_J^2); \quad \alpha_k \sim N(\mu_K, \sigma_K^2); \quad \theta_t \sim N(\mu_T, \sigma_T^2)$$

$$\sum_i \alpha_i = \sum_j \alpha_j = \sum_k \alpha_k = 0$$

We note that the mean of the α terms must be centered at 0 for the identifiability of the model, as written in the third line. The measure of an influencer's "ability" to report correct information about a voter i is given by $\alpha_i + \alpha_j + \alpha_k + \mathbf{X}\beta_X + \mathbf{Z}\beta_Z$, and the average of these predicted values over all of the voters sampled in a polling booth give an estimate of the average ability to report about voters in the polling booth. The θ_t parameter is a parameter that measures the ease of guessing the piece information (the higher, the easier to guess). The parameter vectors β_X and β_Z are used to test the hypotheses listed. We also note that a more general version of the item response model using "discrimination parameters" which adjust for the relative importance of each piece of information in building a measure of ability. However, we require far more pieces information to measure in order to incorporate this form of analysis without strong priors.

4.2.3 Model 3

Two of the variables in reported in the constructed index in model 2 are particularly difficult to guess correctly, the level of education of the voter and the number of rooms in home of the voter. As alluded to in the above discussion, the ability to guess these variables correctly may indicate a particularly high level of knowledge of the voter. In order to test how the influencer does on these difficult to ascertain pieces of information, we consider a generalization of the item response model for these two items.

To develop our model, we break our education variable into sensible categories for the Indian system: less than primary education/illiterate, class 5 passed, class 8 passed, class 10 passed, and class 12 passed. For the room characterization, we consider dwellings in the categories of: 1 room, 2 rooms, 3 rooms, and 4 or more rooms. We model the deviation from the true answer by considering three levels of deviation: 2 or more categories difference, 1 category difference, and correct answer. That is, we create a deviation variable for voter $i \in I$, influencer $j \in J$, and polling booth $k \in K$, Δ_{ijk} , with values $q \in \{0, 1, 2^+\}$. Our space of pieces of information, T , consists of the combined 9 educational and dwelling categories listed above. We can now write down a generalization of the item response model in ordered logit model:

$$P(\Delta_{ijk} \geq q) = \text{logit}^{-1} (c_q - \mathbf{X}_i\beta_X - \mathbf{Z}_i\beta_Z - \alpha_k - \alpha_j - \alpha_i + \theta_t)$$

$$\alpha_i \sim N(0, \sigma_I^2); \quad \alpha_j \sim N(0, \sigma_J^2); \quad \alpha_k \sim N(0, \sigma_K^2); \quad \theta_t \sim N(\mu_T, \sigma_T^2)$$

$$\sum_i \alpha_i = \sum_j \alpha_j = \sum_k \alpha_k = 0$$

$$q \in \{1, 2^+\}; \quad c_{2^+} \leq c_1$$

We again note that the mean of the α terms must be centered at 0 for the identifiability of the model. Again, the measure of an influencer's "ability" to report correct information about a voter i is given by $\alpha_i + \alpha_j + \alpha_k + \mathbf{X}_i\beta_X + \mathbf{Z}_i\beta_Z$, and the average of these predicted values over all of the voters sampled in a polling booth give an estimate of the average ability to report about voters in the polling booth. The parameter vectors β_X and β_Z are used to test the hypotheses listed. The parameters c_{2^+}, c_1 correspond to "cutpoints" between the categories of $2^+/1$, and $1/0$, respectively.

5 Hypotheses and Analyses for Paper 2: Who Are Influencers Willing to Help? (similar to point 2b above)

In a second series of analyses (see point 2b above), we make inferences as to which villagers each of these local actors are willing to help and assist in their interactions with the state. To measure which citizens each influencer is willing to help, we rely on several responses provided by influencers during the Game 1 and Game 2 sections of [INFLUENCER QUESTIONNAIRE \(DAY 2\)](#).

In assessing the willingness to help voters, we consider two types of dependent variables. The first dependent variable is a ranking of willingness to help citizens (1 – 12) of sampled voters in the polling booth area, where 1 is most willing to help. The second dependent variable considers how each influencer allocates 6 tokens over 12 sampled voters in the polling booth. The predictors in the regression models are specified in the second tab of [Variables and Measures declared in PAP](#).

5.1 Hypotheses

- H16.** Influencers are more willing to help citizens from their own caste group than citizens from other groups. (Caste bias)
- H17.** Influencers are more willing to help citizens with whom they share a partisan preference or affiliation than other citizens. (Partisan bias)
- H18.** Influencers are more willing to help poorer and more disadvantaged citizens than other citizens. (Pro-poor bias)
- H19.** influencers are more likely to help citizens whom they have already helped in the past or with whom they are in an existing reciprocal relationship. (Reciprocity bias)

H20. influencers are more likely to help citizens with whom they are closely connected or know well (Connectivity bias)

We also specify hypotheses that relate to the interaction between the identity of influencers (on several dimensions) and this first series of hypotheses.

H21. The connectivity bias should be stronger for T3 influencers than for T1 and T2 influencers.

H22. The pro-poor bias should be stronger for influencers who are currently elected.

H23. the partisan bias should be stronger for strongly partisan influencers.

5.2 Models and Analyses

5.2.1 Model 1

We use the variable R to denote the rank variable. Models of dependent variables that are ranked can be difficult to analyze. Here we construct a model that is appropriate for this particular data and discuss some of the underlying requirements of the data. We again return to an ordered logistic framework, while noting that each number between 1 and 12 (inclusive) will be used once in each polling booth. This creates more structure on the data than the usual model, where a value can be used multiple times by the same actor, as in the models above. Our model evaluates the expected probability of each sampled voter of being the unique person in the sample given a particular rank, using a behavioral assumption. In particular, we assume that if the sampled voters are not differentiated in any way, then they have equal probabilities of being given each rank. Furthermore, since each rank is guaranteed to be given over the sample, the sum of the expected probabilities of being given a rank is necessarily 1, and the expected probability over the sample is necessarily $\frac{1}{12}$.

The fact that the probabilities of each rank are equal, when voters are undifferentiated, implies that the cutpoints in the ordered logistic model are fixed with known values. In particular, given the discussion above, for a rank r , we know that a particular cutpoint, c , will take the value:

$$\text{logit}(P(R \leq r)) = \log\left(\frac{\frac{r}{12}}{1 - \frac{r}{12}}\right) = c$$

It also implies that the expectation at the polling booth level \mathbb{E}_K over some predictor x yields:

$$\mathbb{E}_K(\text{logit}(P(R \leq r))) = c + \mathbb{E}_K(x) = c \Rightarrow \mathbb{E}_K(x) = 0$$

This condition implies that all variables must be mean-centered at the polling booth level. This allows us to write down the model. We will define a matrices, $\tilde{\mathbf{X}}$, which will be understood to be the matrix \mathbf{X} with each column mean-centered at the polling

booth level. This necessarily implies that the influencer level predictors are identically 0, so these must be removed from $\tilde{\mathbf{X}}$. Notice that the constraints at the polling booth level imply that there is no variation in the model at the polling booth or influencer level.

At the same time, it would be inappropriate include no way of controlling for the influencer or the polling booth in the data. The trick is to realize that the coefficients on $\tilde{\mathbf{X}}$ themselves should vary by influencer and polling booth effects. Let \mathbf{W} be a matrix of influencer-level predictors (those that were removed from $\tilde{\mathbf{X}}$), and \mathbf{Z} defined as before. To control for voter, influencer, and polling booth effects further, we consider random effects for voter i in polling booth k , α_{ik} , in addition to the coefficient-wise random effects for influencer j , γ_j , and polling booth k , γ_k . This also addresses the concern that the data are clustered by influencer and polling booth. We may now write down an appropriate ordered logit model:

$$\begin{aligned}
 P(R_{ijk} \leq r) &= \text{logit}^{-1}(c_r - \tilde{\mathbf{X}}\beta_{jk} - \alpha_{ik}) \\
 \beta_{jk} &= \beta_{\tilde{\mathbf{X}}} + \mathbf{W}_i\beta_W + \mathbf{Z}_i\beta_Z + \gamma_j + \gamma_k \\
 \gamma_j &\sim N(\mathbf{0}, \Sigma_J); \quad \gamma_k \sim N(\mathbf{0}, \Sigma_K) \\
 \alpha_{ik} &\sim N(0, \sigma_{IK}^2); \quad \sum_i \alpha_{ik} = 0 \text{ for each } k \in K \\
 r &\in \{1, 2, \dots, 11\}; \quad c_r = \log\left(\frac{\frac{r}{12}}{1 - \frac{r}{12}}\right)
 \end{aligned}$$

The hypotheses of interest are tested from the coefficient vectors $\beta_{\tilde{\mathbf{X}}}, \beta_W, \beta_Z$. We note that the random effect for the voter is centered at 0 at the polling booth level to fit the assumptions of the model. Also note that the random influencer and polling booth effects are modeled as vectors drawn from a multivariate normal distribution with a single variance-covariance matrix. As discussed above, the cutpoints in the model do not have to be estimated as they are defined by the behavioral assumptions in the data structure. Finally, note that since probabilities are estimated for each rank, the predicted values can be expressed as an expected rank.

5.2.2 Model 2

Estimating allocation behavior, and targeting biases, with this data can be difficult. The model developed here extends upon a similar method developed in [Schneider and Sircar \(2015\)](#). There are two major empirical challenges in estimating allocation behavior in this setting. First, the method must account for the fact that the allocator (in this case the influential person) can only allocate a maximum of 6 tokens. Thus, the allocation to potential receivers (in this case the sampled voters) cannot be treated as truly independent. In particular, giving a token to one individual in the population implies that there are fewer tokens to distribute over the rest of the population. In order to rectify this

problem, one has to recognize that the average number of tokens over the population is always identical (the number of tokens divided by the number of voters). If the influential person were randomly choosing recipients for tokens, then each voter would have the identical number of tokens in expectation (the average). Thus, if a voter has a desirable attribute, we expect her to receive a *premium*, an expected number of tokens above the average. The proposed statistical strategy models these premiums, constraining the average number of tokens over the population properly.

This non-independence property across potential receivers applies to desirable attributes of the voters as well. For instance, the number of co-partisans in the population mediates the relative premium in allocation for a co-ethnic receiver by the allocator, i.e., the premium decreases as the number of co-ethnics increases. If the allocator wishes to target only co-ethnics (with no other distinction between individuals) and there are six co-ethnics, the allocator can give one token to each co-ethnic without difficulty. If, however, there are more than six co-ethnics, then it is a certainty that at least one co-ethnic will receive no token. Thus, the relative premium of being a co-ethnic is inversely related to the number of co-ethnics in this situation.

The key observation that allows for identification of the empirical model is that mean allocation in a polling booth is always identical, the number of tokens divided by the number of potential receivers, or $6/12 = 0.5$. This implies that if all the predictors are centered around their means at the polling booth level, the constant term in a regression is fixed. In particular, let y_{ijk} denote the allocation given to voter i by influencer j in polling booth k . Since the number of tokens is in the form of count data, a Poisson regression (accounting for over dispersion) is appropriate. A quasipoisson regression model provides the same mean function as poisson regression, λ_{ijk} , for voter i and influencer j in polling booth k , but allows for over dispersion by estimating variance $\sigma^2 \lambda_{ijk}$ for the observation.¹⁰ However, the salience of each attribute may still vary by influencer and across polling booths (as in the previous model); in order to address this issue, we vary the coefficients by influencer and polling booth. To control for voter, influencer, and polling booth effects further, we consider random effects for voter i in polling booth k , α_{ik} , in addition to the coefficient-wise random effects for influencer j , γ_j , and polling booth k , γ_k .

The complete regression model, using the same matrices defined in the previous model, is then written as:

$$\begin{aligned}
 y_i &\sim \text{Poisson}(\lambda_{ijk}, \sigma^2) \text{ where } \sigma^2 \text{ denotes an overdispersion parameter} \\
 \lambda_{ijk} &= \exp(\beta_0 + \tilde{\mathbf{X}}_i \boldsymbol{\beta}_{jk} + \alpha_{ik}) \\
 \boldsymbol{\beta}_{jk} &= \boldsymbol{\beta}_{\tilde{\mathbf{X}}} + \mathbf{W}_i \boldsymbol{\beta}_W + \mathbf{Z}_i \boldsymbol{\beta}_Z + \gamma_j + \gamma_k \\
 \gamma_j &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_J); \quad \gamma_k \sim N(\mathbf{0}, \boldsymbol{\Sigma}_K)
 \end{aligned}$$

¹⁰In the standard poisson distribution, the variance is fixed at λ_{ijk} , the same as the mean.

$$\alpha_{ik} \sim N(0, \sigma_{IK}^2); \quad \sum_i \alpha_{ik} = 0 \text{ for each } k \in K$$

$$\beta_0 = -\ln 2$$

The hypotheses of interest are tested from the coefficient vectors $\beta_X, \beta_W, \beta_Z$. We note that the random effect for the voter is centered at 0 at the polling booth level to fit the assumptions of the model. Also note that the random influencer and polling booth effects are modeled as vectors drawn from a multivariate normal distribution. As discussed above, β_0 in the model does not have to be estimated as it is defined from the data structure.

6 Hypotheses and Analyses for Paper 3: Which influencers are better at convincing citizens and voters? (similar to point 3 above)

As shown in rows 2-14 of the third tab of the attached spreadsheet titled [List of Variables for PAP](#), we rely on a variety of dependent variables from a variety of instruments in order to measure an influencer's ability to persuade respondents to engage in a costly behavior.

We have several dependent variables in this analysis. In terms of lab game type data, contributions in a dictator game (as percentage of the endowment given) as a function of encouragement corresponding to the treatment groups, as well as the effect on whether or not the citizen contributes in an n -person public goods game.

In terms of preference data, we obtain data as a 4-point scale on whether it is normal for citizens to give time to participatory programs, and whether the respondent would be willing to give 2-3 hours a week to such a program. Again, these dependent variables are to be analyzed by different treatment group encouragements.

The rest of the dependent variables can be characterized as willingness to attend the focus group meeting, and actual attendance at the focus group meeting. In terms of willingness to attend, we collect self-reported data on willingness to attend (taking the values of 0 (no), 1 (maybe), 2 (yes)) from both the initial encouragement and the mobilization encouragement. We also collect binary data on whether the citizen is willing to distribute the two invitation letters.

We then check whether the citizen encouraged attended the meeting, whether another household member attended the meeting, and whether guests receiving invitation letters, and how many, from the citizen attended the meeting, as three separate dependent variables.

Finally, we analyze the influencer's self-reported ability to mobilize and individual (4-point scale), and a ranking over the sampled voters in the polling booth about ability to mobilize. The models for these dependent variables are similar to the first model in section 4 and the first model in section 5, respectively, so we do not write down the models in this section again (the third tab of [Variables and Measures declared in PAP](#) however lists the variables we include in our analyses).

6.1 Hypotheses

We test a number of hypotheses about the ability of different types of influencers to persuade respondents to engage in a costly behavior. The third tab of [Variables and Measures declared in PAP](#) refers to the variable we use to test each of these hypotheses.

- H24.** Citizens who received one of the treatments (T1, T2 or T3) are more likely to be influenced than citizens who did not (i.e. citizens in control group T4)
- H25.** "Real" influencers (T1 and T2) are more likely to influence citizens than "benchmark influencers" (T3).
- H26.** Influencers who are elected or were just elected (until the 2016 elections) in local village institutions are more likely to influence citizens than unelected ones.
- H27.** Influencers who were elected in the past in local village institutions are more likely to influence citizens than unelected ones.
- H28.** influencers who are from a locally dominant group or an upper caste are more likely to influence citizens than influencers from lower castes.
- H29.** Influencers are more likely to influence citizens from their own group than about citizens from other groups.
- H30.** On average, influencers who have been affiliated in the long run with a specific political party are less likely to influence citizens than influencers who have not been affiliated in the long run with a specific political party.
- H31.** Influencers who have been affiliated in the long run with a specific political party however are more likely to influence citizens who have also been affiliated to that party.
- H32.** Influencers who are affiliated with the locally dominant party (i.e. the party that won the seat in the 2015 state elections) are more likely to influence citizens than influencers affiliated with other parties and than influencers affiliated to no specific party.
- H33.** Influencers who have the same partisan preferences than citizens are more likely to influence these citizens than influencers affiliated with other parties and than

influencers affiliated to no specific party.

- H34. Influencers who are assets-wealthy are more likely to influence citizens than influencers who are assets-poor.
- H35. citizens who are assets-wealthy are less likely to be influenced than citizens who are assets-poor.
- H36. Influencers who have a higher level of education are more likely to influence citizens than influencers with a lower level of education.
- H37. citizens who have low levels of education are more likely to be influenced than citizens who are highly educated
- H38. Influencers who are members of non-political organizations or associations are more likely to influence citizens than influencers who are not.
- H39. Influencers whose occupation is non-political but high status or with high-networking potential are more likely to influence citizens than influencers who are not.
- H40. Influencers who are bureaucrats, civil servants or have a state job are more likely to influence citizens than influencers who are not.
- H41. Influencers who have in the past helped a sampled citizen are more likely to influence that citizen than influencers who have not.
- H42. Influencers who have provided a job to a citizen in the past are more likely to influence that citizen than influencers who have not.
- H43. Citizens are more likely to be influenced when they are closely connected to the influencer.

6.2 Models

When analyzing experimental data, it is often advisable to keep the dependent variable untransformed to ease interpretation. As such, we our experimental analysis follows a standard multilevel "least squares" type form. Let $\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4$ be a series of dummy variables denoting whether the sampled voter is in T1, T2, T3, or T4, respectively. Our declared experimental regression controls for clustering at 3 levels, in polling booth $k \in K$, block $b \in B$, and district $a \in A$, and serves as the basis for the power analysis below. Let y be any of the dependent variables (binary or not) in the analysis. Then we estimate:

$$y_i = d_{1i} * \beta_1 + d_{2i} * \beta_2 + d_{3i} * \beta_3 + d_{4i} * \beta_4 + \alpha_k + \alpha_b + \alpha_a + \varepsilon_i$$
$$\alpha_a \sim N(0, \sigma_A^2); \quad \alpha_b \sim N(0, \sigma_B^2); \quad \alpha_k \sim N(0, \sigma_K^2); \quad \varepsilon_i \sim N(0, \sigma^2)$$

Significance of the appropriate treatment group is calculated by simulating the distribution of the coefficients $\beta_1, \beta_2, \beta_3, \beta_4$ (corresponding to means of T1 through T4) and

looking at their differences. Note that the random effects, the α terms, control for the fact that data are clustered at 3 levels.

We also estimate a more complicated model to test the more nuanced hypotheses. In particular we vary the coefficients on all of the predictors (and the constant term) by treatment group, $d \in \{1, 2, 3, 4\}$, in a Bayesian multilevel framework. The main reason to do this is that with a large number of predictors (as we have) can produce very extreme results due to a loss of degrees of freedom. The Bayesian framework tends to produce more believable estimates from a prediction standpoint, and it is worth noting that the random selection of polling booths and voters fit the assumptions of random effects models at those levels. The model is below:

$$\begin{aligned}
 y_i &= c_d + \mathbf{X}_i \boldsymbol{\beta}_{X,d} + \mathbf{Z}_i \boldsymbol{\beta}_{Z,d} + \varepsilon_i \\
 c_d &= c + \alpha_k + \alpha_b + \alpha_a + \alpha_d \\
 \boldsymbol{\beta}_{X,d} &= \boldsymbol{\beta}_X + \mathbf{Z}_i \boldsymbol{\delta}_Z + \gamma_k + \gamma_b + \gamma_a + \gamma_d \\
 \boldsymbol{\beta}_{Z,d} &= \boldsymbol{\beta}_Z + \zeta_k + \zeta_b + \zeta_a + \zeta_d \\
 \begin{pmatrix} \alpha_k \\ \gamma_k \end{pmatrix} &\sim N(\mathbf{0}, \boldsymbol{\Sigma}_K); \quad \begin{pmatrix} \alpha_b \\ \gamma_b \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_B); \quad \begin{pmatrix} \alpha_a \\ \gamma_a \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_A); \quad \begin{pmatrix} \alpha_d \\ \gamma_d \end{pmatrix} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_D) \\
 \varepsilon_i &\sim N(0, \sigma^2)
 \end{aligned}$$

The parameter vectors, $\boldsymbol{\beta}_{X,d}, \boldsymbol{\beta}_{Z,d}$, are used to test the hypotheses.¹¹ More specifically, the distributions of these parameters can be generated from the model, and the hypotheses can be tested as the difference in these parameters over the simulation by looking at 95% simulated (credible) intervals.

6.3 Power Analysis

The proposed study is an experiment with 4 treatment groups (or 2 depending on the perspective). Each of the 4 groups contains 3 of the 12 sampled voters (in the day 2 and day 3 citizen interviews) over 176 polling booths. This means that there are 528 individuals in each of the 4 treatment groups. For the remainder of this analysis we define n to be the sample size in two treatment groups (those that are compared). So, because the treatment groups are envisioned to be the same size, we define $n = 1056$.

¹¹Note that we add additional control variables to Z when modeling the effects of predictors on meeting attendance. As listed in the [List of Variables for PAP](#), we also control for the time set for the meeting, the weather on that day (i.e. whether it rained hard), whether the respondent's profession would have likely prevented him from attending a meeting at that hour, and finally, for the number of influencers who actually attended the meeting. These covariates are listed and described in rows of the third tab in [List of Variables for PAP](#).

For a binary variable a conservative estimate of the standard deviation of a proportion (that is equal to 1), p , is $sd(p) = \sqrt{p(1-p)} \leq \frac{1}{2}$. Using this information, for two treatment groups, A and B , a conservative estimate of the standard error of the difference in proportions (assuming no clustering) is given by:

$$se(\hat{p}_A - \hat{p}_B) \leq \frac{1}{2} \sqrt{\frac{1}{\frac{n}{2}} + \frac{1}{\frac{n}{2}}} = \frac{1}{\sqrt{n}}$$

A nearly identical calculation can be implemented for non-binary case. Let σ_y be an upper bound on the standard deviation of y in any treatment group. Then a conservative estimate of the difference in expected values for two treatment groups, A and B , is just:

$$se(\widehat{\mathbb{E}}_A(y) - \widehat{\mathbb{E}}_B(y)) \leq \sigma_y \sqrt{\frac{1}{\frac{n}{2}} + \frac{1}{\frac{n}{2}}} = \frac{2\sigma_y}{\sqrt{n}}$$

For our binary predictors, we seek to assess whether we can detect an effect of 0.1 in the proportion influenced/mobilized. In order to generate the industry standard of 80% power with 95% confidence intervals, we aim to select a σ (the standard deviation of the estimator) such that $\hat{p}_A - \hat{p}_B - 2.84\sigma \geq 0$. Plugging in 0.1 for the difference in means, we arrive at $\sigma = 0.0352$.

There is one final concern in the power analysis, the design effect. The design effect is the inflation factor from a "complex design" applied to the variance from resulting from simple randomization in the experimental design. In short, we account for the fact that our data are cluster-randomized and that this causes inefficiency. The design effect inflates the variance by approximately $1 + (m - 1)\rho$, where m is sample size within the cluster and ρ is the proportion of the variance due to across cluster variance. Our cluster size is 3, since three voters are in each treatment group in each polling booth.

For our analysis, we let $\rho = 0.15$ (usually we think of $\rho \leq 0.1$ as low), which would a reasonably high level of correlation for binary variables (see Chakraborty et al. (2009) on this point).¹² In order to calculate the required sample size, we solve for:

$$(1 + (m - 1)\rho) \frac{1}{n} \leq \sigma^2$$

Using our assumptions, and solving for n , we arrive at $n = 1048.5 < 1056$. This implies that a sample size of 528 per treatment group (over 4 treatment groups) is sufficient to detect an effect 0.1 between any two groups under given assumptions.

We also note that we will run a "bundled" treatment analysis that does not distinguish between real influencers, and the control and benchmark influencers. In this scenario,

¹²In theory, there are at least 3 levels at which variance inflation occurs in our data, at the polling booth level, at the block level, and at the district level. The assumption then is that the combined inflation is less than the upper bound inflation factor we use here. We believe this is defensible because our pilots have found little similarity in mobilization/influence behavior across polling booths.

we have 2 treatment groups, 1 with real influencers, and one without, and the same calculations show that we can detect an effect of 0.07 between the two groups.

Repeating the analysis for the non-binary dependent variables show that we can detect an effect of $0.1 * (2\sigma_y)$ with the 4 treatment group analysis and an effect of $0.07 * (2\sigma_y)$ in the 2 treatment group analysis.

Because all of non-binary variables have a hard minimum and maximum bound, we should expect these effect sizes (0.1 and 0.07) are reasonable when the non-binary variable, y , is re-scaled into a new variable, y^* , to be between 0 and 1. To see why, create a variable y^* by subtracting the minimum bound from each observation and dividing by the maximum bound. Note that y^* is bounded between 0 and 1. For the expected value of y^* , \bar{y}^* , create another variable \tilde{y} that takes the value 0 if $y^* \leq \bar{y}^*$ and 1 if $y^* > \bar{y}^*$. Notice that \tilde{y} is binary and necessarily has higher variance than y^* . Our assertion follows from the fact that our power analysis was calculated as a conservative estimate of a binary variable.

References

- Auerbach, Adam. 2013. "Demanding Development: Democracy, Community Governance, and Public Goods Provision in India's Urban Slums." Manuscript.
- Banerjee, Abhijit, Arun G. Chandrasekhar, Esther Duflo and Matthew O. Jackson. 2013. "The Diffusion of Microfinance." *Science* 341(6144).
- Berenschot, Ward. 2014. Political Fixers and India's Patronage Democracy. In *Patronage as Politics in South Asia*, ed. Anastasia Piliavsky. Cambridge, UK: Cambridge University Press.
- Björkman, Lisa. 2014. "You can't buy a vote: Meanings of money in a Mumbai election." *American Ethnologist* 41(4):617–634.
- Bussell, Jennifer. 2014. "Representation Between the Votes: Informal Citizen-State Relations in India." Unpublished.
- Chandra, Kanchan. 2004. *Why Ethnic Parties Succeed: Patronage and Ethnic Headcounts in India*. Cambridge University Press.
- Chauchard, Simon. 2016. "Why Provide Electoral Handouts? Theory and Microlevel Evidence From Mumbai." Manuscript.
- David A Kim, Alison R. Hwong, Derek Stafford, D. Alex Hughes, A. James O'Malley, James H. Fowler and Nicholas A. Christakis. 2015. "A Randomised Controlled Trial of Social Network Targeting to Maximise Population Behaviour Change." *Lancet* 386(9989):145–153.

- DellaVigna, Stefano and Ethan Kaplan. 2007. "The Fox News Effect: Media Bias and Voting." *The Quarterly Journal of Economics* 122(3):1187–1234.
- DellaVigna, Stefano and Matthew Gentzkow. 2010. "Persuasion: Empirical Evidence." *Annual Review of Economics*, 2010, Vol.2 2(643-669).
- Dewan, Torun, Macartan Humphreys and Daniel Rubenson. 2014. "The Elements of Political Persuasion: Content, Charisma and Cue." *The Economic Journal* 124(574):257–292.
- Dunning, Thad and Janhavi Nilekani. 2013. "Ethnic Quotas and Political Mobilization: Caste, Parties, and Distribution in Indian Village Councils." *American Political Science Review* 107(1):35–56.
- Enos, Ryan D. and Eitan D. Hersh. 2015. "Campaign Perceptions of Electoral Closeness: Uncertainty, Fear and Over-Confidence." *British Journal of Political Science* pp. 1–19.
- Gentzkow, Matthew. 2006. "Television and Voter Turnout." *Quarterly Journal of Economics* 121(3).
- Gerber, Alan S., James G. Gimpel, Donald P. Green and Daron R. Shaw. 2007. "The influence of television and radio advertising on candidate evaluations: results from a large scale randomized experiment." Working Paper, Yale University.
- Green, Donald P. and Alan S. Gerber. 2008. *Get Out the Vote! How to Increase Voter Turnout*. Washington DC: Brookings Intitute Press.
- Hicken, Alan. 2009. *Building Party Systems in Developing Democracies*. Cambridge, UK: Cambridge University Press.
- Holland, Alisha C. and Brian Palmer-Rubin. 2015. "Beyond the Machine: Clientelist Brokers and Interest Organizations in Latin America." *Comparative Political Studies* 48(9):1186–1223.
- Humphreys, Macartan, William A. Masters and Martin E. Sandbu. 2007. "The role of leaders in democratic deliberations: results from a field experiment in São Tomé and Príncipe." *World Politics* 58(4):583–622.
- Kitschelt, Herbert and Steven Wilkinson. 2007. Citizen–Politician Linkages: An Introduction. In *Patrons, Clients, and Policies*, ed. Herbert Kitschelt and Steven Wilkinson. Cambridge University Press.
- Krishna, Anirudh. 2002. *Active Social Capital*. Columbia University Press.
- Kruks-Wisner, Gabrielle. 2015. "Claiming the State: Citizens' Mobility & Demand for Public Services in Rural India." Book Manuscript.
- Larreguy, Horacio, Cesar Montiel and Pablo Querubin. 2016. "Political Brokers: Partisans or Agents? Evidence from the Mexican Teacher's Union." Manuscript.

- Magaloni, Beatriz. 2006. *Voting for Autocracy: Hegemonic Party Survival and its Demise in Mexico*. Cambridge, UK: Cambridge University Press.
- Manor, James. 2000. "Small-Time Political Fixers in India's States: 'Towel over Armpit'." *Asian Survey* 40(5):816–835.
- Schneider, Mark. 2014. "Does Clientelism Work? A Test of Guessability in India." Unpublished.
- Schneider, Mark and Neelanjan Sircar. 2015. "Whose Side Are You On? Identifying Distributive Preferences of Local Politicians in India." CASI Working Paper, 15-01.
- Sircar, Neealanjan. 2015. "A Tale of Two Villages: Kinship Networks and Political Preference Change in Rural India." CASI Working Paper, 15-02.
- Srinivas, M. N. 1955. The Social Structure of a Mysore Village. In *Village India: studies in the little community*, ed. McKim Marriott. University of Chicago Press pp. 1–35.
- Stokes, Susan C., Thad Dunning, Marcelo Nazareno and Valeria Brusco. 2013. *Brokers, Voters, and Clientelism: The Puzzle of Distributive Politics*. Cambridge, UK: Cambridge University Press.
- Van de Walle, Nicholas. 2007. Meet the new boss, same as the old boss? The evolution of political clientelism in Africa. In *Patrons, Clients, and Policies*, ed. Herbert Kitschelt and Steven Wilkinson. Cambridge University Press.
- Weiner, Myron. 1967. *Party Building in a New Nation: The Indian National Congress*. University of Chicago Press.