THREE ESSAYS IN DEVELOPMENT ECONOMICS

BY

CHITRA JOGANI

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Economics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2019

Urbana, Illinois

Doctoral Committee:

Associate Professor Rebecca Thornton, Chair
Associate Professor Richard Akresh
Assistant Professor Tatyana Deryugina
Assistant Professor Benjamin Marx

Abstract

This dissertation comprises of three chapters on understanding the effectiveness of public policies in a democracy. Chapter 1 titled "Effect of Political Quotas on Candidate Attributes and the Provision of Public Goods" studies the effect of such quotas on attributes of political candidates and on the provision of public goods. I use a regression discontinuity design that exploits the assignment of caste quotas in the latest redistricting in India. I find quotas lead to political candidates with lower wealth, lower criminal records, but similar education levels. The difference in attributes is also observed and is more pronounced for the stronger candidates: those affiliated with political parties, and those elected for office. The caste quotas also increased the representation of women in politics. I find no significant difference in the level of public goods currently available in rural India between quota-bound and non-quota-bound areas. The results suggest an increase in political diversity with no negative effects on the provision of basic facilities.

Chapter 2 titled "Does More Schooling Infrastructure Affect Literacy?" studies how the expansion in schooling infrastructure affects the female literacy rate using the Education for All program in India. I exploit the variation in the targeting of the program to educationally and not educationally backward sub-districts. Using regression discontinuity and panel data of all schools in India, I find that there was a significant expansion in the total number of schools, number of girls' schools, and residential schools for girls, in the educationally backward areas. But being classified as educationally backward did not lead to a significant effect on either the female literacy rate or the gender gap in literacy rate. To achieve a quicker solution to low levels of literacy, alternative cost effective methods compared to large scale infrastructure programs can be explored.

Finally, Chapter 3 titled "Spatial Analysis of an Education Program and Literacy in India" explores the presence of spatial dependency for studying the association between an education program and literacy. To do this, I use data from a nation-wide education program in India which involved building schools and increasing the accessibility of education to girls. The program targeted geographical districts that were educationally backward or had a low rate of rural female literacy. The paper finds significant spatial correlation in the educational backwardness of districts and in the outcome of interest, change in rural female literacy rate. To account for the spatial dependency, I fit a spatially distributed error model (SDEM). The SDEM estimates suggest a 0.08

percentage point increase in the rural female literacy rate and a 0.02 percentage point decrease in the gender gap in literacy rate with a one point increase in the educational backwardness of a district. The SDEM estimates are similar to the estimates from a spatially blind model, and there is no additional influence of the program received by the neighboring districts on the change in rural female literacy rate of a district.

To my Father

Table of Contents

Chapter 1

EFFECT OF POLITICAL QUOTAS ON CANDIDATE ATTRIBUTES AND THE PROVISION OF PUBLIC GOODS

1.1 Introduction

To address the underrepresentation of women and people belonging to certain ethnicities, castes, and other minority groups in politics, policymakers often turn to quotas. Quotas increase the representation of political candidates from the underrepresented identity. But these policies remain controversial because of fears that the quality of political candidates could fall, and economic development could decrease.[1] To address this controversy, in this paper I study the following questions: whether quotas lead to political candidates with different attributes, and whether the presence of quotas affect the provision of public goods.

To assess the effect of quotas, I use political quotas which have been in place for the past seven decades in the largest democracy, India. The quota exists for the historically disadvantaged groups, the Scheduled Castes and Scheduled Tribes, who comprise a quarter of India's population. I use quotas in state elections where approximately a quarter of the total 4,120 electoral districts in India are "reserved".[2] Reservation of an electoral district for Scheduled Castes (Tribes) stipulates only citizens belonging to the Scheduled Castes (Tribes) can stand for elections from the district. The winner from the election in a district is the elected official or the district's representative in the state legislative assembly. Reservation status thus guarantees representation of these castes and tribes in influential positions, as members of the legislative body of the state government.

Estimating the causal effect of quotas can be challenging as the assignment of quotas are not

[1] Affirmative action policies in college admissions or employment raise similar concerns; that they may lead to candidates or employees of lower qualifications or ability (Bagde et al., 2016; Holzer and Neumark, 1999). Affirmative action policies are controversial because, by stipulating that minorities must be represented, members of other groups have fewer opportunities to gain admission to prestigious higher education institutions or to assume leadership positions within organizations, for example.

[2] In India, members of the central government are selected in national elections; the representatives of the state legislature are elected through state elections; and representatives at the lower level of government are elected through local elections.

random, leading to endogeneity issues. The presence of quotas can be correlated with other characteristics of electoral districts, such as being economically poor. But, the assignment of reservation status to constituencies by the Delimitation Commission in India provides a suitable empirical setting for tackling the endogeneity problem. The reservation status of an electoral district (or constituency) depends on the population share of the reserved groups. Although there is no explicit population cutoff, I exploit the procedure of reservation in a novel way to establish a discontinuous relationship between the share of the reserved population and the reservation status of constituencies. I implement a regression discontinuity design to estimate the causal effect of quota.

The primary findings on how the use of quotas affects the attributes of candidates for office: Candidates standing for election from scheduled caste constituencies are less likely to be criminals (4.4 percentage points); they have lower assets (0.14 million USD or 76 percent lower); and they have similar education levels to candidates in constituencies not reserved for the scheduled castes.[3] Comparing the estimates with statistics on attributes for the overall population, the results suggest that the difference observed in the attributes of political candidates do not merely reflect differences in the attributes of the populations of reserved and unreserved castes. Second, in scheduled caste constituencies, more women seek election (5 percentage points more than in non scheduled caste constituencies), and more women win (8 percentage points). Thus, caste quotas not only increased political representation from the targeted population, but also women.

I also explore whether the results hold for candidates who differ in their party affiliation, incumbency status, and winning status. I find the difference in attributes across reservation status of constituencies is higher among candidates affiliated with political parties compared to candidates who contest independently (35 percent of candidates are independents). The differences in attributes is also observed for the non-incumbents, who may not have the advantage that are typically associated with incumbency. The effect is more pronounced among the strongest candidates, the winners, who are also the elected official. The phenomenon of candidates with criminal records also winning elections can be because other potential candidates for the same office also have criminal records, or because voters prefer some other attribute of the candidate, such as caste. Using data from a voters opinion survey, I do find citizens from unreserved castes and tribes care more about the caste or religion of the candidate than citizens from the reserved groups do. Thus, the possibility of caste-biased voting in the unreserved constituencies can lead to selection and winning of criminal candidates (Banerjee and Pande, 2007; Besley et al., 2005).

[3]As a measure for desirable attributes, such as honesty or competence, I use information on criminal charges, education level, and wealth of candidates.

To explore whether areas that are bound by quotas have lower provision of public goods, I use the data on facilities in all villages in 2011. The results do not indicate a significant difference caused by reservation status in the availability of facilities, such as schools, hospitals, roads, and banks across constituencies. The size of the estimates imply, any effect greater than a decrease in the availability of facilities in 4 percent of villages (or three to four villages) in a reserved constituency can be ruled out. There is also no evidence of a difference in the growth of facilities for the reserved constituencies in the 2001-2011 period. This is in line with findings of similar insignificant effects of caste quotas on economic development for the period of 1971-2001 (Jensenius, 2015). Thus, the reserved constituencies are on par with similar constituencies that are unreserved, at least in availability of the basic facilities.

This paper relates to several literature. First, this paper contributes to the literature on the effect of affirmative action policies on quality of candidates. Most studies in the affirmative action literature have focussed on the effect of policies in college admissions and employment opportunities (Holzer and Neumark, 1999; Bagde et al., 2016). This paper studies how political quotas causally affect attributes of politicians who are in a crucial position of managing a state. Although the attributes that define a "good" politician is not clear, to the extent that having criminal charges can be considered a bad attribute and the level of education a good attribute, then the reserved constituencies are better off in this respect.[4] Another issue in the affirmative action literature is that affirmative action policies targeting one minority group may displace people from other underrepresented groups (Bertrand et al., 2010). But, this paper suggests that may not always be the case, as caste-based political quotas in fact led to more representation of the other underrepresented group in politics, women.

Second, this paper also builds on the literature on importance of identity (such as gender, caste, or ethnicity) and characteristics of a politician. Several studies have used quotas to estimate the effect of the identity of a political leader (Chattopadhyay and Duflo, 2004b,a; Bardhan et al., 2010; Dunning and Nilekani, 2013). Other papers have pointed out that the effectivenss of a leader may also depend on his other attributes, such as honesty, integrity, and ability (Besley, 2005; Besley et al., 2005). In addition, recent studies have emphasised the phenomenon of selection of candidates on the basis of character (Bernheim and Kartik, 2014). This paper also suggests evidence of selection of citizens who seek elections and also selection of candidates by parties on the basis of

[4]Existing evidence imply election of criminal politicians affect development negatively (Prakash et al., 2014; Chemin, 2012), they under utilize development funds and who have lower attendance rates in meetings (Gehring et al., 2015). Educated leaders can contribute to higher growth (Besley et al., 2011) but may not affect education outcomes (Lahoti and Sahoo, 2017).

the candidate's attributes.

Finally, this paper adds to the few studies on the effect of political quotas in state legislatures on provision of public goods (Jensenius, 2015; Min and Uppal, 2011). Various studies have focussed on quotas in the local village council where the assignment of quotas is randomized, unlike the quotas in the state legislature. In a large and decentralised system of government, as in India, the state legislature plays a crucial role, such as being responsible for law making, providing resources to the local government, making government schemes available to citizens, and to use the constituency development fund for ensuring development in the constituencies. The closest paper to this paper is Jensenius (2015). The paper uses propensity score method to find no impact of reservation for Scheduled Castes on several development indicators in the period of 1971-2001. Using new districts with data after the latest redistricting, I find a similar null result for a wider range of village facilities by implementing a regression discontinuity design. I find an insignificant result on reservation for Scheduled Tribes as well.[5]

The rest of the paper proceeds as follows: Section 1.2 provides some background on political representation in India and the process of reservation of constituencies. Section 1.3 describes the theoretical expectations and conceptual framework for the effect of quotas. Section 1.4 explains the research design and data. Section 1.5 presents and discusses the results on the effect of quotas. Section 1.6 concludes.

1.2 Institutional Background

1.2.1 Quotas and Elections in India

Quotas are a form of mandated political representation for underrepresented populations.[6] Political quotas in the state elections in India exist for the Scheduled Castes and Scheduled Tribes,

[5]The literature has found different effect of quotas for Scheduled Castes and Scheduled Tribes. For example, Chin and Prakash (2011) find no impact on poverty when the number of assembly constituencies reserved for Scheduled Castes in a state increases, but do find a decrease in poverty on increasing the share of seats reserved for Scheduled Tribe. Whereas, Krishnan (2007) find leaders from Scheduled Castes improve primary schooling facilities, but finds no significant effect for leaders from the Scheduled Tribes. Pande (2003) finds a positive effect of reservation for Scheduled Tribes on welfare spending, and a positive effect of reservation for Scheduled Castes on job quotas.

[6]There can be various kinds of quotas based on their nature of restriction (candidate list vis-a-vis reserved seats). See (Bird, 2014; Htun, 2004; UNDP, 2012) for details and case studies on political quotas.

which comprise a quarter of the population (16.6 percent and 8.6 percent, Census of India 2011). Such castes and tribes have been historically disadvantaged, with people, sweepers and cobblers, for instance, treated as lower caste or untouchables. The lower castes have faced discrimination and exploitation by the upper castes for generations. By contrast, tribal communities traditionally resided in forest areas, which led to their geographical and cultural isolation. Having such a history of oppression, members of such castes and tribes may lack the confidence to voice their opinion or to seek a political career. It would also be difficult for them to compete with candidates from the upper castes. Therefore, as a measure of positive discrimination, the constitution implemented political quotas for the Scheduled Castes and Tribes after the independence of India (1947).[7]

Quotas in the state legislative assembly, which exists in the form of reserving electoral districts for the Scheduled Castes and Tribes, guarantee seats for them in the state legislative assembly. The members of the state legislative assembly (known as MLA) are elected in state elections, which occur every five years. The state elections use a "first-past-the-post" system; several candidates run for office in an electoral district, and the candidate with the highest number of votes is the winner or the MLA. There are a total of 4,120 electoral districts or assembly constituencies; thus elections from the 4,120 districts lead to 4,120 MLAs. Reservation of an electoral district for Scheduled castes (or Tribes) mandates that only candidates belonging to the Scheduled castes (or Tribes) are allowed to run for office from that district. However, voting within reserved districts takes place in the same way voting takes place in all districts; that is, voting is open to everyone, not just to those from reserved groups. Approximately a quarter of the 4,120 districts are reserved; hence reservation means that a quarter of the MLAs in India are represented by members of scheduled castes and tribes.

Seats in the state legislative assembly increase the representation of scheduled castes and tribes in an influential position. The state legislature has significant power over law making, and many matters related to issues, such as agriculture, local governments, and police. It also participates with the central government in decisions related to matters such as education, and marriage. The members of the state legislative assemblies participate in the legislative meetings and make important decisions for the state. They are responsible for ensuring development in their constituencies, suggesting projects for implementation to bureaucrats, and for providing access to different govern-

[7]Quotas for the Scheduled castes and tribes also exist in the local and national government. In state and national elections electoral districts are reserved, whereas for local elections it is implemented as reservation of seats in the local council. There also exists quota for women in the local elections, but the proposal of quotas for women in state and national government is still under discussion. Several other affirmative action policies for the Scheduled castes and Tribes exist in the education and employment sectors in India. To address atrocities against the lower castes, special courts were established under the Prevention of Atrocities Act of 1989 (Girard, 2016).

mental schemes to people. They have complete access to the constituency development fund, the MLA-Local Area Development fund (MLA-LAD). Additionally, they can nominate members for other bodies, such as block development committees (Wilkinson, 2006) and they have the power to transfer bureaucrats (Iyer and Mani, 2011; Nath, 2015).[8]

1.2.2 Process of Redistricting and Reservation in India

The Delimitation Commission defines the boundaries of electoral districts (or constituencies) during redistricting. Redistricting or delimitation divides states into equally populous constituencies using data from the latest census, and is supposed to occur every 10 years.[9] The first redistricting took place in 1953, followed by the second and third in 1961 and 1971. But, there was no redistricting in the period 1971-2007. Evidence suggests that the freeze in redistricting was not because of any political manipulation (Bhavnani, 2015). The reason for the freeze was to not punish states achieving lower population growth with lower representation in the state and national government. Hence, the latest redistricting in 2007 occurred after a gap of three decades. Figure A.1 shows the latest redistricting led to a significant change in the boundaries of constituencies. [10]

The Commssion also decides the reservation status of constituencies during redistricing. Figure 1.1 presents the distribution of constituencies based on their reservation status. Assembly constituencies are classified according to their reservation status into three groups: unreserved or General (GEN), reserved for the Scheduled Castes (SC), and reserved for the Scheduled Tribes (ST). A constituency that is either a SC or a ST constituency is a reserved constituency in Figure 1.1.

The Delimitation Commission uses the following procedure to determine the allocation of reserved constituencies to states and the reservation status of each constituency. The total number of constituencies reserved is proportional to the population share of the reserved group in India. Likewise, the number of reserved constituencies in a state is proportional to the population share

[8]In this paper, I focus on quotas in the state legislative assembly. I also confirm some results for quotas in the national elections, which elect members for the lower house of India's Parliament or Lok Sabha.

[9]For the national elections, the country is divided into 543 electoral districts known as parliamentary constituencies. A parliamentary constituency is composed of several assembly constituencies. An assembly constituency always lies completely within a parliamentary constituency.

[10]Sources of maps for Figure A.1: Old constituencies from Sandip Sukhtankar and Manasa Patnam, New constituencies from Devdatta Tengshe of Datameet. The states of Arunachal Pradesh, Assam, Manipur, Nagaland, Jharkhand, Jammu and Kashmir were not delimited in 2007.

of the reserved group in the state. Thus, a state with a higher fraction of the reserved population will have a higher proportion of reserved constituencies.

To maintain geographic heterogeneity of reservation for scheduled castes, there is an extra step of allocation of SC constituencies across administrative districts.[11] The number of SC constituencies (or seats) entitled to a district is equal to the total number of SC constituencies allocated to the state multiplied by the relative population share of scheduled castes in the district. But, this number can be a fraction and is thus called the predicted number of SC constituencies for the district. Figure 1.2 shows the rule followed by the commission for final allocation of SC constituencies across districts, which has to be an integer, and is dtermined as a step function of the predicted number of SC constituencies for the district. Following the allocation of constituencies, the next step is to determine the reservation status of the constituencies. To do this, the Delimitation Commission reserves constituencies with the highest population share of Scheduled Castes in the district, and highest population share of Scheduled Tribes in the state. I explain the procedure in the empirical section and using examples in section A.2.1 of the data appendix. The reservation procedure forms the basis of my empirical strategy.

1.3 Theories and Conceptual Framework

1.3.1 Potential Impact of Political Quotas

Whether political quotas serve the purpose of benefiting the minorities or worsen the situation for the entire constituency remains controversial. One of the direct effects of quotas is that they guarantee politicians of a particular identity, such as gender or caste. Several studies have evaluated the importance of identity of the politician for election or economic outcomes. For example, the gender of the candidate is believed to have favorable development outcomes in interests of the representative gender or the population (Chattopadhyay and Duflo, 2004b; Clots-Figueras, 2011, 2012; Iyer et al., 2012). But, reserving the position of chief in the local government for women led to a decrease in targeting of resources towards other underrepresented groups (Bardhan et al., 2010). Under caste and tribal quotas at the local level, where the chief of the local government belonged to one of these minorities, the evidence has been mixed; from weak distributive effects (Dunning and Nilekani, 2013) to postive benefits (Bardhan et al., 2010; Chattopadhyay and Duflo,

[11] The Scheduled Tribes population is concentrated in some states and is approximately half of the Scheduled Castes population.

2004a). The religion of the political candidate has also been found to be important in influencing health and education outcomes (Bhalotra et al., 2014).[12] These studies are based on citizen-candidate models where the identity of the politician might influence his or her policy position or policy preference.

But, it has also been argued that the identity of the politician does not matter if political party influence is higher (Jensenius, 2015). Likewise, it has been argued that because politicians care about their own interests and careers, even elected officials who are members of the representative castes or tribes are unlikely to pay special attention to their own groups because their incentive is to try to please the majority population and, thus, the voter pool needed for them to remain in power.

Fear of quotas leading to negative effects is based on several hypotheses, such as that quotas could result in candidates who are ill-suited for the responsibilities of being a leader. Such candidates would have less bargaining power, be less effective in attracting resources for their constituencies. Such a situation would result in a worse allocation of resources to all the people in the constituency and would affect development. Another hypothesis put forward is that candidates in a reserved constituency could experience lower competition because people from the unreserved categories are ruled out from standing for elections. The leader from a reserved constituency can also take his powers for granted because the presence of the quota gives him a greater chance of remaining in power.

Similar concerns have been raised about affirmative action policies in education or employment. Critics have argued that such policies might lead to admission of students or hiring of employees who are ill prepared for the position, and this would be detrimental to their careers. However, Holzer and Neumark (1999) find that employees hired under affirmative action had lower educational qualifications but not lower performance. Using affirmative action policy in engineering colleges in India, Bagde et al. (2016) did not find any evidence of a mismatch between students and colleges. In case of political quotas, this question is important, as having people who are incompetent to be politicians may not only be detrimental for their own career but also for citizens

[12]Such identity of the politician could potentially influence outcomes through many channels. Examples of such channels include higher complaints by women in presence of women leader about goods they prefer more (Chattopadhyay and Duflo, 2004b), and bargaining power of the legislator (Pande, 2003). Similarly, having MLAs from the same community can lead to a decrease in the cost of complaining for people from these communities, either directly or through local officials. MLAs have access to funds for development, and are in a position to discuss issues that require attention in the state legislative meetings. Furthermore, MLAs themselves can be ministers of different departments, such as health, railway, and education, putting them in greater positions of power.

of the state.

Apart from identities such as gender, caste, or ethnicity that a person is born with, there has been recent emphasis on character of the politician. As mentioned in (Besley, 2005; Akerlof and Kranton, 2000) a politician has many characteristics that identify him and determine his quality. Characteristics such as criminality and level of education of politicians has also been observed to affect development (Prakash et al., 2014; Chemin, 2012; Gehring et al., 2015; Besley et al., 2011; Lahoti and Sahoo, 2017). Character of the politician is important for political selection of the candidate as well (Besley et al., 2005; Bernheim and Kartik, 2014). However, the characteristics that are essential for a "good" politician might be difficult to define. These can be subjective characteristics such as charisma, personality, intelligence, integrity; or objective measures such as education, income, experience (Murray, 2015). In addition, quality of the candidate who is finally in office might be affected by the institutional setting and the method of political selection (Besley, 2005). To explore if quotas could lead to politicians of different charcteristics, I present a conceptual framework in the next section.

1.3.2 Conceptual Framework

The process of election in India and the final selection of a candidate in office can be understood using the following stages:

Stage 1: Some citizens of India decide to run for political office from a constituency. The decision of a citizen to run depends on the cost (c_i) and return (R_i) from running.

Stage 2: Parties nominate candidates of type t_i, where t_i=$t(x_1, x_2, x_3, ..)$, a function of different attributes of the candidate $x_1, x_2, ..$ and others. Some examples of the attributes $x_1, x_2, ..$ are wealth, popularity, ability or education level. A party may require some minimum qualification for a candidate to be eligible, $t_i >= \underline{t}$. A candidate has a value of $V(t_i)^w$ on winning and $V(t_i)^l$ on losing to a party, where the probability of win for the candidate is $p(t_i)$. For the purpose of generality, I am not assuming any restriction on the values of $V(t_i)^l$, and there can be situations in which the values are negative, positive or zero. Simialrly, not making assumptions regarding the relation between $V(t_i)^w$ and $V(t_i)^l$, or for value of the candidate to the party when $t_i = \underline{t}$.

While selecting candidates, the parties would internalize preferences of the voters, but might also have its own set of desirable characteristics, such as loyalty to the party (Besley, 2005). The $\underline{t}$

could be institutional restrictions for contesting elections, but also minimum criteria that a political party may have for the different attributes of a candidate or the $x's$. Once a candidate meets the minimum criteria, the party would want to select a candidate with the maximum expected value, $E(V(t_i))$, where $E(V(t_i)) = p(t_i)V(t_i)^w + (1 - p(t_i))V(t_i)^l$.

Stage 3: Several candidates contest from different political parties in a constituency and voters vote for their preferred candidate.

Finally, the candidate with highest vote is the winner or the Member of Legislative Assembly (MLA) who is responsible for functioning of the state government and for development in the constituency. This framework can also be generalized to other democracies that have a similar process of election.

Given that these stages would determine the outcome of elections for constituencies, then quotas could influence the mechanism of an election in one or more of the stages, and we can expect that this could lead to politician with different attributes in office. With the implementation of quotas, the difference lies in the fact that some constituencies gain a status of being restricted to citizens only belonging to certain castes or tribes. Several studies have assumed that the cost of contesting or the benefit from winning is different for people from the underrepresented group (Chattopadhyay and Duflo, 2004b; Besley et al., 2005). Such restriction could alter the incentive of the agents involved in the above stages. For example, reservation could decrease the cost of contesting elections (c_i), or increase the return from contesting elections (R_i) for a candidate from the reserved group.

Similarly, the decision to contest may depend on being selected by politicial parties. During selection of candidates by parties, if belonging to the Scheduled Castes or Tribes is one of the x or the attribute for the type of a candidate, then this does not remain an unconstrained parameter in the reserved constituencies anymore. Caste of the candidate can be an attribute for selection by parties because it can also be a determinant of voter's preference. For example, voters could prefer candidates belonging to their own caste. Parties may internalize such preferences during their selection of candidates.

Therefore, whether candidates of similar attributes will emerge from a reserved or general constituency will depend on the distribution of t_i for the population and the reserved groups, or on the distribution of t_i for citizens who want to stand for elections in Stage 1. The selection will also depend on the selection of candidates by parties from the reserved and unreserved constituencies

in Stage 2, and ultimately the candidates chosen by voters in Stage 3. Even in absence of quotas, the type of candidate selected could be different based on other factors, such as the population composition of the reserved or unreserved constituencies. Hence, to causally estimate the effect of reservation on otherwise similar constituencies, I use the empirical framework described in the next subsection.

1.4 Empirical Framework

In this section I describe the empirical strategy used to identify the causal effect of quota. The challenge in estimating the effect of quotas is the issue of endogeneity, since the reserved constituencies differ from the unreserved constituencies in characteristics other than the reservation status. But, the procedure of redistricting and assignment of reservation status to constituencies in India provides a quasi-natural experiment setting. I exploit this to set up a regression discontinuity design (RD). I also describe the data sources used for the exercise and confirm the RD assumptions.

1.4.1 Empirical Strategy: Regression Discontinuity

The Delimitation Commission reserves constituencies based on the population share of the reserved group. Using the procedure followed by the Commission, I am able to establish a discontinuous relation between the reservation status and population share of the reserved group in the constituency. To achieve this, I rank constituencies based on the population share of Scheduled Tribes within a state and Scheduled Castes within a district in a descending order. A rank of one implies highest population share of the reserved group. The number of constituencies reserved in the district for these castes (and in the state for these tribes) acts as the cutoff rank. Hence, constituencies with rank less than or equal to the cutoff have a reservation status of one.[13] Figure 1.3 shows this for reservation of SC constituencies. Instead of using the discrete variable rank of a constituency, I use the continuous variable population share of Scheduled Castes as the assignment variable, and the population share of Scheduled Castes in the last constituency reserved (one with the cutoff rank) as the cutoff. Hence, all constituencies with percentage of Scheduled Castes population higher than the cutoff have a reservation status of one. I normalize the cutoff to zero. All other points are differences of the population share of Scheduled Castes from the cutoff, which I refer to as the deviation of the percentage of Scheduled Castes population. I follow a similar

[13] I explain the procedure in the form of an algorithm in section A.2.1 of the data appendix.

procedure for reservation of ST constituencies.

Figure 1.4 presents the relationship between the reservation of a constituency for Scheduled Tribes and the normalized population share of Scheduled Castes. The figure shows the probability of reservation of a constituency for Scheduled Castes increases by 0.95 on crossing the cutoff and not one. This arises due to a few exceptional rules. For example, a constituency may be eligible for reservation for both Scheduled Castes and Scheduled Tribes due to a high relative population share of the groups. In such cases, the constituency is reserved for Scheduled Tribes.[14] Additionally, to distribute SC constituencies over the state, the Commission avoids spatially contiguous constituencies for reservation of Scheduled Castes. On observing the map of assembly constituencies, this seems to be an explanation for some cases. To illustrate, in the state of Andhra Pradesh, the constituency Addanki although eligible for reservation, did not receive reservation status. This decision was made because Addanki is adjacent to Santhanuthalapadu (shown in Figure A.2), which has the highest percentage of Scheduled Caste populations in the Prakasam district. As a result, Santhanuthalapadu was reserved, Addanki was skipped, and Yerragondapalem was reserved instead.[15]

Figure 1.4 also shows that there are far more unreserved constituencies compared to reserved constituencies, leading to fewer points on the right. Also, there are fewer constituencies with extreme percentages of Scheduled Castes populations. Table 1.1 provides the first stage estimates for SC reservation and the estimate obtained is 0.95. The estimate remains similar under various specifications, and choice of bandwidths. As shown in Figure 1.5, on crossing the threshold for Scheduled Tribes population, the probability of a constituency being reserved for Scheduled Tribes jumps from zero to one.

[14] An example of this is the constituency of Habibpur in the state of West Bengal. It has the highest Scheduled Castes population in the district of Maldaha, but because of its high scheduled tribe population in the state, the district received reservation for such tribes. Another situation is when a district is not assigned any reserved seat for scheduled castes, even though, according to the rule, it must receive one. This can happen if the total number of seats a state should receive is less than the sum of the individual entitlements of the districts. For example, based on the population share of Schedule Castes relative to other states, the state of Haryana has an allocation of 17 scheduled caste seats. But, the number of districts in the state is 19. Based on the rule for assignment within districts, each of the 18 districts should receive one scheduled caste constituency. Thus, the district of Mahendragarh with the lowest scheduled caste population share does not receive any such seat. I have excluded districts with no scheduled caste seats; hence, these districts will not cause the fuzziness we see in the RD. But if a constituency was not reserved for similar reasons when a district should receive more than one reserved AC, it was included.

[15] Yerragondapalem had the second highest percentage of Scheduled Caste populations in the Prakasam district among the unreserved constituencies, and constituency Parchur had the highest. But, Parchur was spatially adjacent to Santhanuthalapadu as well.

I am interested in estimating the effect of reservation on the outcome variables of interest: attributes of political candidates and the level of public goods in the constituency.

Figure 1.5 shows the reservation status for ST constituencies changed deterministically at the population cutoff. Thus, the treatment effect for ST reservation can be estimated using a sharp regression discontinuity design. The regression specification can be represented by the following equations:

$$Y_i = \alpha + \beta D_i + f(X_i) + \varepsilon_i \tag{1.1}$$

$$D_i = \begin{cases} 1 \text{ if } X_i >= 0 \\ 0 \text{ if } X_i < 0 \end{cases} \tag{1.2}$$

where D_i=1 implies constituency i is reserved for Scheduled Tribes, X_i: Difference in percentage of Scheduled Tribes population from the cutoff, Y_i is an outcome variable of interest.

But for SC reservation, D_i is not a deterministic function of X_i. There is a discontinuous change in probability of the treatment status (D_i) at the cutoff, that is

$$\lim_{x \uparrow c} Pr[D_i = 1 | X = x] \neq \lim_{x \downarrow c} Pr[D_i = 0 | X = x] \tag{1.3}$$

but the change in the probability of treatment is less than one. Thus, the treatment effect for SC reservation is estimated using a fuzzy RD design. Estimation of a fuzzy RD is similar to the two stage least-squares method as shown by the following equations:

$$Y_i = \alpha + \beta D_i + f(X_i) + \varepsilon_i \tag{1.4}$$

$$D_i = \alpha_1 + \beta_1 Z_i + g(X_i) + \mu_i \tag{1.5}$$

$$Z_i = 1(X_i >= 0) \tag{1.6}$$

where D_i=1 implies constituency i is reserved for Scheduled Castes, X_i is the difference in percentage of Scheduled Castes population from the cutoff, where the cutoff has been normalised to be zero. Y_i is the percentage of villages in the constituency with the public good or the different attributes of a candidate. Z_i: dummy variable which takes the value of one to the right of the cutoff

and zero to the left of the cutoff. Z_i is used as the instrument for the endogenous variable D_i.

I follow a nonparametric method of estimating the causal effect at the cutoff using local linear regression (Lee and Lemieux, 2010). The coefficient of interest β for the fuzzy design can then be estimated by considering only observations close to the cutoff (c) as below:

$$\beta = \frac{\lim_{x\uparrow c} E[Y|X=x] - \lim_{x\downarrow c} E[Y|X=x]}{\lim_{x\uparrow c} E[D|X=x] - \lim_{x\downarrow c} E[D|X=x]} \tag{1.7}$$

Because the probability of being reserved for Scheduled Tribes jumps from zero to one at the threshold, the jump in the outcome variable is the average treatment effect, that is

$$\beta = \lim_{x\uparrow c} E[Y|X=x] - \lim_{x\downarrow c} E[Y|X=x] \tag{1.8}$$

To balance the tradeoff between bias and precision of the estimates, I consider observations within the Calonico et al. (2014) optimal bandwidth (CCT). I report the bias-corrected robust estimates that measure the average treatment effect at the threshold. For the main analysis of the paper, I focus on constituencies reserved for Scheduled Castes.

1.4.2 Data and Summary Statistics

Redistricting data: I use the delimitation reports for each state to construct a dataset with the population and reservation status of constituencies. These reports from the Election Commission provide an accurate measure of population for the constituency. They contain the total population, the population of Scheduled Castes and Scheduled Tribes, the reservation status, and the administrative area of each constituency. For the analysis, I exclude the following observations from the sample: states that were not delimited in 2008, union territories in India that do not have legislative assemblies, and Delhi (which is a union territory but also the capital of India). This yields 3,397 constituencies (of the 4,120 possible constituencies) in 21 states. The current identification strategy is made possible by the availability of these delimitation data and can be used for future research.[16]

[16]The population of the constituencies for the earlier delimitations is not made publicly available. The other source of population is the Census of India. However, it is difficult to use the Census to obtain constituency-level populations because the Census provides population data only for administrative divisions.

Data on Affidavits: To study the effect of quotas on the attributes of candidates, I use data on the candidate affidavits. In the wake of a 2002 Supreme Court ruling, all candidates seeking election to political office in India must file personal affidavits. The affidavits contain information on the gender, education level, criminal charges, and the assets and liabilities of the candidates.[17] I complement this with data on state elections in India. These contain names of political candidates from different assembly constituencies and the total number of votes they received for all state elections from 1977 to 2016.[18]

Table 1.2 summarizes the data on affidavits.[19] The table shows that political candidates from the reserved and unreserved constituencies differ significantly on most attributes. Candidates from general or unreserved constituencies have approximately thrice the assets compared to candidates from the reserved constituencies. The number of criminal charges including serious crimes is higher on average for candidates from general constituencies. Representation of women in all constituencies in state elections is strikingly low; Males represent 90 percent of candidates from reserved constituencies, and 94 percent of the candidates from unreserved constituencies. Fewer contestants view for office in reserved constituencies. There is no significant difference in the education level of candidates from the two types of constituencies.

Voters' Survey: To understand voters' preference about politicians and public facilities, I use the the Daksh voter participation survey of 2014. The survey has 238,694 respondents. These data are unique for both the sample size and the extensive set of questions used.[20] The respondents provided their opinions on the performance of the leader, and the factors that influence their vot-

[17]Source: ADR affidavit data crawled from Myneta.com and cleansed by Trivedi Centre for Political Data (TCPD). I thank Gilles Verniers, Rajkamal Singh, and the TCPD team for providing the data. I add details on incumbents and perform some other checks from the Myneta.com website. Details on the filing of affidavits and their veracity can be found in Bhavnani (2012); Vaishnav (2012); Prakash et al. (2014)

[18]The source of the data for the years 1977 to 2012 is Bhavnani, Rikhil R., 2014, India National and State Election Dataset", http://dx.doi.org/10.7910/DVN/26526 Harvard Dataverse Network [Distributor] V1 [Version]. I use these data to construct the margin in elections, to document other election outcomes of those constituencies, and to conduct various cross checks. For post-2012, I use various web-based sources.

[19]Out of the sample of 3,397 constituencies, the data could be matched for 3,378 constituencies. I have considered affidavits of candidates for state elections from 2009 to 2014 for all the states in my sample that had at least one election after the redistricting. Recent data are available for some states that had a second round of elections after 2014; these data, however, do not have all the variables included in the previous data. The estimates for the common variables remain similar when including the recent data.

[20]Daksh India conducted the survey, and I obtained the data from Datameet, a community of Data Science and Open data enthusiasts. The survey was regarding the Member of Parliaments (MPs) instead of MLAs. Nevertheless, the survey provides evidence to some extent about the perception of politicians.

ing decisions. Panel A of Table 1.3 shows the candidate characteristics that survey respondents consider to be important. I represent the data by whether the respondent belong to the reserved groups. People from unreserved groups consider the caste or religion of the candidate more important compared to respondents from the reserved groups. People from general caste also seemed to value distribution of gifts by the candidate which is illegal.

Panel B of Table 1.3 summarizes information on how voters ranked the importance of different public goods and services, such as education, health, and agriculture. Participants indicated whether they considered certain services to be of High, Low or Medium importance.[21]The table shows that approximately half of the respondents thought these were of high importance. The difference between the opinion of people from the reserved and unreserved group was significant, especially given the large sample size. More people from the Scheduled Castes and Tribes considered basic facilities of education, health, agriculture, and electricity to be of high importance in comparison to people from general castes. The table shows the system of job reservation which exists for the Scheduled Castes and Tribes in India, was of lesser importance to people from the general caste, who do not benefit from it.

Census Data: As a measure of public good provision, I use the 2001 and 2011 Census data, which contain information on facilities for all villages in India. This exercise required linking several datasets. This process was not straightforward. First, mapping administrative divisions to constituencies in India has been a challenge in the past, and several studies have followed different approximation methods.[22] Some papers have recently used mapping between villages and the old constituencies (Jensenius, 2015; Asher and Novosad, 2017). But, with new boundaries of constituencies, the villages have to be mapped to the new constituencies.[23] Second, the Census 2001 and 2011 data use different village codes. After several rounds of matching and cleaning, the final sample of 2,801 constituencies is used for the analysis on public goods.

[21]The Daksh voter perception survey included many questions under each category. For example, under the category of agriculture, the questions addressed the importance of agriculture loans, prices of agriculture commodities, irrigation, subsidies for fertilizers, and so on. To summarize the information, I create a dummy indicating "High importance" for each variable and find the average for the entire category. I followed the same procedure for the other categories.

[22]For example, Blakeslee (2013) aggregates the values at the subdistrict level and maps these to Parliamentary constituencies.

[23]I sincerely thank Raphael Susewind for sharing the 2011 mapping between villages, and the new constituencies. The data are protected under the Open Data Commons Open Database license. I followed several procedures to check the consistency and correctness of the data, comparing them with other publications of the Delimitation Commission. See section A.2.2 of data appendix for details on the cleaning and modifications of the data.

The Census data provides information on facilities available to the population living in the villages of India. The data contain information on whether a village has a facility, such as a primary school, a middle school, or a health center. I define aggregate variables based on the different facilities. For example, the variable "Middle School or higher" in Table 1.4 is a dummy variable that takes the value of one if a village has a middle school, or a secondary school, or a senior secondary school; otherwise, the value is zero. I follow a similar strategy for other variables. I then aggregate the data at the constituency level to find the percentage of villages in a constituency with a facility.

Table 1.4 compares the average level of village facilities by reservation status of the constituencies; 57.6 percent of the villages in a general constituency have a middle school or above and 36.2 percent of the villages have a health center or a hospital.[24] The reserved constituencies have a lower level of the public goods, but the difference between a general and a scheduled caste constituency is small. There has been convergence in the level of facilities, and the gap between the general and reserved constituencies has reduced in comparison to data from earlier Censuses (Banerjee and Somanathan, 2007; Blakeslee, 2013).

I also investigate the growth in the facilities for the 2001-2011 period by constructing a panel of villages. Doing so was difficult because the boundaries of constituencies had changed due to the latest redistricting. Many villages now belonged to a different constituency, meaning that the new constituencies formed had different composition of villages than those in 2001. To solve this, I calculate the level of public goods for the new constituencies in 2001 with their current composition of villages.[25]

1.4.3 Regression Discontinuity Assumptions

The RD analysis is valid under certain assumptions: there must be no manipulation of the treatment variables around the cutoff; the covariates are balanced across the cutoff; and the assignment variable is continuous (Lee and Lemieux, 2010). I provide evidence to support that each of these assumptions is satisfied. The results are provided in the appendix.

[24] In the 2011 Census, almost all villages (approximately 95 percent) report having a primary school after the government's effort to ensure universal primary education in India.

[25] This provides a hypothetical estimate of what would have been the level of public goods in these constituencies if they had existed in their current form in 2001. I only consider variables that were present in both the Censuses, and aggregate the variables in a way to make them comparable. This is an approximation because some of the villages that are now mapped to the new constituencies experienced a different administration in 2001.

Manipulation of the treatment variable would imply changing the percentage of the Scheduled Castes or Scheduled Tribes population relative to the cutoff to affect reservation status for some constituencies. This is unlikely for several reasons. Firstly, the process of delimitation happens after the population Census has been recorded and published. Redistricting is performed by the Delimitation Commission which has no connection to the Census Division of India. The population numbers are used for many other purposes apart from reservation of constituencies, thus suspecting that they will be changed for the purpose of determining reservation status is farfetched.

Second, manipulating the variable for the population share of a scheduled caste or tribe would be difficult, requiring perfect manipulation of the reserved population or the total population. Moreover, accomplishing this would also require manipulating the variable relative to the population of other constituencies to address the relative ranking of constituencies and, hence, their reservation status. Another way in which there could be manipulation would be through manipulating the electoral boundaries or gerrymandering. This too seems unlikely since the Delimitation Commission responsible for the delimitation is an independent organization comprised of members without any political connection or affiliation. The latest redistricting was mostly politically neutral (Iyer and Reddy, 2013). To draw the boundaries for constituencies in a district, the commission proceeds in a zig zag manner starting from the north, proceeding northwest and then turning south. Additionally, the shape of constituencies as seen in Figure 1.1 reduces suspicions about gerrymandering.

The second assumption requires that constituencies do not differ in other characteristics discontinuously around the cutoff. I test for whether covariates (such as population size, population of other castes, average number of households, and the facilities) were balanced in the pre-period using data from the Census of 2001. Figure A.3 in appendix presents the discontinuity plots for the covariates. The plots do not show any significant discontinuities around the cutoff for any of the variables. The final assumption of the assignment variable being continuous holds true because the percentage of the population of scheduled castes and tribes are continuous in nature.

1.5 Empirical Results

1.5.1 Effect of Reservation on Characteristic of Candidates

Figure 1.6 presents the RD plot for the attributes of candidates based on affidavits declared before the state elections. The figure shows that candidates running for office from SC constituencies

have a lower level of total assets in comparison to candidates from non-SC constituencies. Candidates from SC constituencies are less likely to be criminals, and they have fewer numbers of serious criminal charges against them. There does not seem to be a significant difference in the level of education between the candidates from these constituencies. The plot for the number of candidates shows there are fewer political candidates seeking office from SC constituencies.[26]

Table 1.5 presents the RD estimates. The assets are in millions of rupees (1 million rupees = 15,000 US dollars). It is clear that SC reservation causes candidates with lower total assets to stand for elections; assets are 76 percent lower, representing a magnitude of 9.7 million rupees (0.14 million USD).[27] Candidates from reserved constituencies are less likely to have a criminal record (4.4 percentage points), 0.16 lower number of criminal cases and 0.13 lower number of serious criminal charges (33 percent and 43 percent lower compared to the mean of the control). The estimates imply that there is no difference in the level of education among political candidates from reserved or unreserved constituencies, whereas, the levels of literacy and education are lower for the scheduled caste population. On average there also seems to be two fewer candidates contesting from SC constituencies, but the number of females contesting from the reserved constituencies is higher by 5 percentage points. [28]

Hence, restricting only people from Scheduled Castes to stand for elections from the SC constituencies led to candidates with lower criminal charges, lower assets, and increased the representation of females. One of the aim of affirmative action policies has been to increase diversity in the sector of implementation (Epple et al., 2008). This seems to have been achieved to some extent here. Not only do we have an increase in representation of people from the scheduled castes and tribes, but also women. I also perform the analysis for candidates in national elections, in which case Parliamentary constituencies are reserved. I follow similar strategy of using the algorithm for reserving Parliamentary constituencies, and use an instrumental variables strategy to identify the effect. The analysis, provided in section A.1.1 of the appendix, shows the same relationship

[26] I present the analysis for SC reservation only. The results for ST reservation remains similar but due to a smaller number of ST constituencies, the estimation is imprecise (results available upon request). Also, here the comparison is between constituencies reserved and not reserved for Scheduled Castes (but can be reserved for Scheduled Tribes).

[27] The coefficient obtained on estimating the regression for logarithm of the total assets is -.57 which translates to .76 on using $\exp^{beta} - 1$. There were around 850 cases in the entire sample with total assets reported as zero. The coefficients change slightly if I substitute for zero total assets with a value of one, or a number between zero and one, before taking the logarithm. The estimates in Table 1.5 exclude candidates with very high assets; candidates in the top .1 percentile.

[28] For Table 1.5, I consider the bandwidth of equal length on the left and right of cutoff. However there are far more observations from unreserved constituencies than reserved. The estimates do not change if I consider unequal bandwidths (for example smaller bandwidth for to the left of the cutoff).

between reservation and attributes of candidates.

To investigate any heterogeneity in the results, I analyse the different subsamples of candidates based on their party affiliation, whether they were winners, and their incumbency status.

Party Affiliation: Candidates can seek election in India either independently or through affiliation with a political party. To account for any difference between independent and party-affiliated candidates, I estimate the regression for the samples separately. Candidates may choose independent status if they were not selected by a party, or if they chose not to join a party because they did not find a party that aligns with their interests. Candidates with party affiliation from SC constituencies have lower asset holdings (13.17 million rupees/0.2 million U.S. dollars lower, on average); they are less likely (by 6 percentage points) to have a criminal record than party affiliated candidates from non-SC constituencies. The estimates remain negative for the sample that includes only independent candidates; estimates are lower by approximately 50 percent and insignificant for the asset and criminality variables, although imprecisely estimated. The estimates however are negative and significant for the number of males and number of candidates. Thus, the difference is more pronounced for candidates that are selected by parties. The difference is also significantly higher for candidates selected by major parties.[29]

The data suggests some correlation between party candidates and having criminal charges or higher level of assets.[30] The fact that candidates with party affiliation have higher assets has also been observed in the past for major parties in India (Duraisamy and Jérôme, 2017; Vaishnav, 2012; Bhavnani, 2012). Additionally, criminality and assets of a candidate have also been found to be highly correlated (Vaishnav, 2012; Dutta, 2015). Candidates with high assets are strongly preferred by parties because such candidates can provide additional funding to parties for elections or for any emergency (Dutta, 2015; Besley, 2005; Mukhopadhyay, 2014). Although voters do seem to penalize candidates with criminal records (Banerjee et al., 2011, 2014), parties may select candidates with a criminal past to intimidate the voters from opposing parties (Aidt et al., 2011).

Incumbents and Winners: Incumbents could have a different probability of winning or value to the party, such as they can be expected to have some incumbency advantage.[31] Additionally,

[29]Results for independent candidates and major parties are available upon request.

[30]Results available upon request.

[31]There has been mixed evidence regarding incumbency advantages in India. Some papers have found that incumbency provided an advantage before 1991 but a disadvantage later (Linden, 2004; Anagol and Fujiwara, 2016; Uppal, 2009). However, recent research shows that after mandating declaration of the affidavits in 2002, also led to incumbency advantage because worse candidates chose not to run for office (Fisman et al., 2017). Thus, the long

assets of a politician could grow as he is in office (Bhavnani, 2012) and thus incumbents from constituencies can be expected to have higher assets. Thus, I also perform the analysis for only the non-incumbents who would not have an incumbency advantage. The table shows that the results hold for non incumbents as well.

Winners by definition can be considered to be the strongest candidate in both the reserved and unreserved constituencies. To explore whether the difference is observed for even the strongest candidates from both constituencies, I perform the analysis for winners only. With respect to winners, who are also the elected official or the MLA, the effect is much larger. Winners generally have higher assets than other candidates on average, which holds true even if we exclude incumbents. On average, winners have four times the assets of losers, the difference being higher for general candidates than Scheduled Castes candidates. The regression for difference in attributes only for the losers would lead to lower estimates. For example, for asset holdings, the estimate for winners is approximately 5.5 times of that of the losers.[32] There is a correlation between the amount of assets and the status of winning for a candidate. But, the questions whether a candidate's assets are a predictor, and the broad questions surrounding the factors that may determine winnability are issues for future research.

One of the reasons for obtaining the above results could be that people from the reserved groups are more honest or have lower wealth in general, and therefore candidates from these groups have the same attributes too. But, that does not seem to be the case. This is also intuitive as it is unlikely that politicians are a random draw from the population. Unfortunately lack of data prevents verifying this hypothesis by determining the level of criminal charges or asset holdings for the entire population. I instead provide some other statistics to infer the characteristics of the population. For example, 3.7 percent of Scheduled Castes older than 15 years have an educational qualification of "graduate and above"; the comparable figure for people belonging to non-reserved categories is 10.6 percent.[33] But, there is no difference in the level of education of the political candidates. Similarly, the rate of rural poverty is 31.5 percent for the Scheduled Castes population and 22.7 percent for those in non-reserved categories. Thus, the rate of rural poverty is 38 percent higher

established incumbency disadvantage of politicians in India may be reversing now as winners of last election may not want to contest again if they have low chance of winning and were involved in corruption as they have to declare their assets if they decide to re contest.

[32]Results available upon request.

[33]Source: NSS report on Employment and Unemployment, 2011-2012. The numbers are reported using the classification of SC, ST, Other Backward Castes (OBCs), and "others" in these reports. The percentage of graduates and above: from the SC (3.7 percent), ST (3.1 percent), OBCs (6.2 percent) and others (15 percent). People from the "others" category have the highest educational qualifications.

for members of the Scheduled Castes populations. The magnitude of the difference between urban poverty is similar. This suggests that the magnitude of difference we observe among political candidates from reserved and unreserved constituencies is not driven solely by the difference in characteristics of the population.[34]

The results suggest strategic selection by parties in Stage 2 of the conceptual framework, since the estimate is significantly larger for party-affiliated candidates than for independents. One major difference between general and reserved constituencies is that in the reserved constituencies all the candidates belong to the Scheduled Caste; hence, the competition is not on the basis of Scheduled Caste status, and there is no possibility of choosing someone from the other unreserved castes. However, in the general constituencies, there might be a preference of politicians from the same caste or higher caste. So, parties speculate whether they can win on the basis of the caste of the candidate even if the candidate is a criminal. Also, a party is more likely to nominate criminal candidates if they have an electoral advantage in constituencies where the party faces strong competition (Dutta, 2015; Mukhopadhyay, 2014). Strategic nomination might also encourage a party to nominate criminal candidates if other opposing parties are doing so; party leaders may believe that they need someone equally "powerful"; or because criminality could be effectively neutralized as a dimension for voters to make their choices if candidates of opposing parties had criminal records.

Finally, following the overall pool of candidates have different attributes, the estimates for winners in Stage 3 show a similar result. Selection of and winning by criminal candidates in India has received significant attention. Nevertheless, how or why candidates with serious criminal charges win elections has remained an interesting question. There can be different scenarios under which criminal candidates manage to win elections. First, voters may not be aware of criminal charges against a politician; if informed they would not prefer such candidates (Banerjee et al., 2011, 2014; Ferraz and Finan, 2008). Second, the voter may be aware of criminal charges but does not consider them to be true, or voters perceive it as normal for political candidates to have such charges, and the charges are of no consequence to them. Alternatively, criminal charges may signal that the candidate is someone powerful, and thus capable of protecting the citizens of the constituency from other criminals (and politicians) (Vaishnav, 2012). Furthermore, voters may not be able to judge a candidate on the basis of criminality if other strong candidates also have criminal charges.

In open or unreserved constituencies voters may choose candidates with criminal backgrounds if they prefer other characteristics of the candidate, such as the candidate's caste. Using data from the

[34] Although, in a general constituency, people from all castes can seek election, candidates from the higher castes dominate (Pande, 2003; Nath, 2015).

voter perception survey, I find that people from the unreserved castes cared more about the caste or religion of the candidate; in all, 20 percent of the respondents from the general caste considered the caste or religion of the candidate to be very important; this compares to 9 percent of respondents from the reserved castes. Such caste-based voting can lead to the election of people of lower quality (Banerjee and Pande, 2007). It is possible that, because people in reserved constituencies effectively cannot vote based on caste, they may be interested in finding out other attributes, or they may care more about the other attributes of a person as they cannot vote based on caste. By contrast, in unreserved constituencies, the median voter may have more preference of selecting someone from high caste and not make the effort to be informed about other attributes of a person.

1.5.2 Effect of Reservation on Provision of Public goods

Figure 1.7 presents the RD plots for the presence of schooling, health, and transportation facilities, and for electricity availability based on data from the 2011 census.[35] The plots show that there is no significant difference in the level of facilities between SC and non-SC constituencies that are close to the cutoff. The RD estimates for the variables are provided in Table 1.6.[36] The top row represents estimates without state fixed effects, whereas the bottom row represents estimates on including state fixed effects. The estimates are small, and they decrease with the inclusion of state fixed effects. The estimates become negative for all the three variables, but remain insignificant. The standard errors have been clustered at the district level, and the null result obtained is precise.

The largest effect can be observed for the variable "Middle School or higher" with state fixed effects. The estimates imply that on being reserved for SC, percentage of villages that have a middle school or higher decreases by 2.3 percentage points. Considering the standard errors, any effect larger than a decrease of 4.2 percentage points can be ruled out. The average number of villages in a SC constituency is 150, and thus we can expect three to four villages to be affected. This would amount to a maximum of 5,000 people to be affected.[37]

[35] I have restricted my analysis to districts with at least one SC constituency.

[36] The bandwidth considered has been implemented using Calonico et al. (2014), and considers the bandwidth of equal length on the left and right of cutoff. This leads to more observations (approximately double) on the left than the right as there is a higher number of general constituencies compared to the reserved. The regressions can be specified to have a smaller bandwidth on the left such that the observations are approximately equal. This can lead to a change in the sign of the coefficient, but the results would remain within the confidence interval of the initial estimation. However, this procedure would increase the standard errors.

[37] Considering that 70 percent of India's population is rural, this amounts to 0.83 billion people in 600,000 (0.6 million) villages. Therefore, one can estimate that, on average, 1,383 people reside in a village. This implies a

The RD analyses for credit, recreation, drinking water, and communication facilities are presented in Figure 1.8. Although there is no significant difference in the facilities, the graph shows a decrease in the availability of credit facilities as one moves towards constituencies with higher percentage of Scheduled Castes population. Table 1.7 provides the estimates for the above variables; the largest effect is observed for the variable "Credit facilities". The null result obtained is precise, and any effect greater than 4.6 percentage points can be ruled out. The results obtained for ST reserved constituencies using the sharp regression discontinuity design also lead to insignificant results. The results are provided in figures A.4 and A.5 in the appendix. However, the smaller number of scheduled tribe-reserved constituencies results in larger standard errors.

Number of Village Facilities: Next, I examine if any change occurred in the intensive margin of the village facilities; whether the number of such facilties in the villages changed. The census provides information on the number of village facilities for all the items under each category, such as number of middle schools, number of secondary schools, and so on. But, aggregating these variables to have a total number for the entire category is difficult because of possibilities of double counting. I perform the analysis for the disaggregate variables; the results do not change with regard to the discontinuity.[38]

Growth in the Village Facilities: Finally, I test whether any change occurred in the growth of the facilities. I use the panel data of village facilities for 2001 and 2011. Figure A.6 of appendix provides results. The figure shows that while there has been an increase in facilities in the 2001-2011 period, the reserved constituencies have not gained differentially compared to the unreserved. As mentioned in section 1.4.2, the level of public goods in 2001 for the new constituencies is an approximation because some villages in new constituencies were under different constituencies, and experienced different administration in 2001. In an ideal scenario the analysis would be to compare constituencies that were hundred percent similar, in terms of composition of villages and reservation status. However, for constituencies that did not face a huge change in their composition of villages (that is, they did not experience a major change in their boundaries) or a change in reservation status, this should be a reasonable method of approximation. To understand size of the change in the boundaries before and after the delimitation, I overlay the maps of the old and new constituencies.[39] Despite significant boundary changes for many constituencies, the median

maximum of 5,000 people living in three to four villages.

[38]Results available upon request. I have also performed the analysis for different methods of aggregation and the results remain the same.

[39]Details of the analysis and explanation of the overlap percentage can be found in section A.1.2 of appendix.

overlap percentage between the old and new constituencies is 60 percent.

The regression discontinuity results indicate that the level of village facilities in a reserved constituency is similar to the level in a comparable unreserved constituency in 2011. It is important to note that the result is for the overall effect of a quota or reservation status. The main channel through which quotas can affect provision of public goods is through the politician or leader. But, there can be other channels, too, through which quota might affect development. For example, the central government might want to direct more resources towards the reserved population, in which case they could direct resources to constituencies labeled as reserved. Usually the effects of quotas have been used interchangeably with the effects of the leader; however, it is difficult to rule out offsetting heterogeneous effects or complementaries between these channels.

There can be several possibilities for the null result. First, based on the leader channel, and in line with previous speculation, it is possible that the influence of the party is higher, and that the MLA is taking actions based on party decisions. Moreover, the intentions of a leader from the reserved or unreserved constituencies may not be very different; both leaders might cater to people from the reserved communities and provide the basic facilities to gain their votes. But, in the alternative scenario - that is, in absence of reservation - there is a higher chance of having a politician from the unreserved groups, even in the current reserved constituencies. Thus, most likely there will be lower representation of leaders from the reserved groups. Also, if we speculate that leaders from reserved groups are puppets in the hands of the party, it is difficult to know if leaders from unreserved groups in a similar constituency would avoid being puppets in the hands of the party, or rule without fear of losing due to a weak opposition. Second, this is the situation as of 2011, and convergence in these constituencies with respect to the facilities in villages has been observed. Such convergence can be a result of several targeted programs by the Central Government of India (involving building schools and health centres), because areas with higher percentage of minority have also been poor and backward.

1.5.3 Robustness Checks

A possible caveat in the analysis for public goods is that the new reservation status was effective in state elections after the redistricting in 2007. Some states held elections only a year prior to the 2011 census, and, thus, the new reservation status had been in effect for a shorter time. To address this, I rerun the analysis by restricting the sample to states that had elections at least two years before the 2011 census. The results are presented in Table A.1 in appendix. There is no change in the

nature of results obtained. Some areas have been in a reserved constituency for a longer time than others. However, the boundary changes of constituencies make it difficult to perfectly control for the duration of reservation status of constituencies. Nevertheless, these changes should not affect the nature of the results to a great extent. Moreover, Jensenius (2015) obtains similar insignificant results on development in observing the effect of the reservation status over a longer time horizon of three decades, during which no change in boundaries or reservation status took place.

To address the fact that aggregation of village-level facilities might average out effects faced by individual villages, I examine the change in facilities for individual villages that were affected by the latest redistricting. For this, I use the change in boundaries due to the redistricting as an exogenous shock. This led to villages changing constituencies and, in some cases, also reservation status. The analysis for the same is provided in section A.1.3 in the appendix. I find similar insignificant results. Additionally, earlier papers in the literature have analyzed the effect of reservation of constituencies at the district level, that is, how having an assembly constituency reserved affects districts. Section A.1.4 in the appendix provides discussion of carrying out a similar analysis in the recent setting.

1.6 Conclusion

Around a hundred of countries across the world use political quotas to guarantee representation for the minorities in politics. This study examines the effect of quotas on the attributes of political candidates, and on the provision of public goods. Using latest data from India and regression discontinuity, I find that candidates from reserved constituencies (bound by quotas) differ in characteristics from the unreserved regions (not bound by quotas). In particular, the system of political quotas has given rise to a selection of candidates who have lower financial assets, and who are less likely to have criminal records. Education levels of candidates are similar, regardless of whether quotas are in place. Quotas designed to ensure the representation of scheduled castes and tribes have increased the representation of women, even though this was not the stated intent. There is also no significant difference in the level of public goods currently available in rural India between constituencies that are reserved and not reserved.

It is worth mentioning that there might be other unmeasured or psychological gains of having political leaders from reserved categories of castes and tribes. Such candidates may act as a role model, and they make people from the reserved groups more comfortable in approaching politi-

cal authorities, who could also perhaps understand their problems better. The fraction of political candidates and winners from the reserved groups and women is significantly low in the unreserved constituencies. This may imply it is unlikely for people from the minorities to gain representation in absence of quotas.

Quotas in the form of mandated political representation continue to exist in India, and there have been several demands for extending them to people from other categories. Understanding the current relevance and different impacts of quotas that were implemented several decades ago (since 1951) would help in creating and revising effective policies. Additionally, understanding the defining attributes of a "good" politician remains an open question. Precise knowledge of the desired attributes could help in determining appropriate eligibility requirements for political candidates. Policies to increase voter awareness regarding the characteristics of candidates could lead to candidates with undesirable characteristics losing their electoral advantage. Furthermore, knowledge of the complementariness between different attributes and performance of a politician might help to better understand the different channels of influence of various institutional policies.

1.7 Tables

Table 1.1: First stage estimates for SC Reservation

VARIABLES	(1) Reserved for SC
RD estimate	0.941***
	(0.0150)
Observations	1,667
Bandwidth	CCT
Control	% of SC Population

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The number of observations is the number of constituencies within the optimal CCT bandwidth (5.1). The standard errors are clustered at the district level. The estimates remain similar for different selection of bandwidths.

Table 1.2: Attributes of Candidates by Reservation Status of Constituencies

	GEN	SC	ST
Total Assets	21.57	6.68	7.51
Movable Assets	7.10	1.54	2.20
Immovable Assets	14.47	5.14	5.32
Liabilities	3.39	0.72	0.74
Net Wealth	18.18	5.96	6.77
No. of Criminal cases	0.50	0.26	0.25
No. of Serious Crimes	0.30	0.16	0.18
Has a Criminal Record	0.19	0.13	0.12
Males	0.94	0.90	0.90
Age	44.68	44.57	44.89
No. of Candidates	13.44	11.43	8.37
College and Above	0.40	0.40	0.38
N	29,183	4,785	2,018

Information for candidates contesting state elections post 2008. The asset holdings are provided in million of Indian rupees (1 million rupees = 15,000 US dollars) and illustrates the average amount of assets held by a candidate from a constituency type.

Table 1.3: Voters Opinion in India by Caste of the Respondent

	GEN	SC	ST
Panel A: Importance of characteristics of candidates			
Candidate	0.369	0.444	0.413
Party	0.245	0.249	0.298
Caste/Religion	0.206	0.091	0.109
PM candidate	0.222	0.185	0.188
Gifts distribution	0.098	0.052	0.063
Panel B: Importance of facilities to the voters			
Electricity	0.426	0.453	0.470
Health	0.446	0.486	0.502
Agriculture	0.401	0.419	0.436
Education	0.431	0.475	0.495
Transport	0.460	0.474	0.488
Job Reservation	0.394	0.416	0.429
Employment	0.474	0.474	0.485
Defence/Safety	0.394	0.395	0.403
N	81,818	41,325	18,985

The table summarizes the responses of individuals in the Daksh Voter Perception survey by the caste of the respondent. Panel A represents the percentage of respondents who think the mentioned characteristic of the candidate to be very important, whereas Panel B represents the percentage of respondents who considered the indicated facility to be very important. The survey was based on a large sample and the preferences are statistically different across people from different caste.

Table 1.4: Level of Village Facilities by Reservation Status (2011)

Variables	GEN	SC	ST
Middle School or higher	0.576	0.554	0.499
Hospitals/Health Centres	0.362	0.328	0.314
Transport	0.626	0.575	0.473
Electricity	0.851	0.838	0.656
Credit facilities	0.313	0.270	0.152
Tap	0.643	0.595	0.495
Recreation facilities	0.518	0.536	0.471
Phone/Post Office	0.919	0.915	0.846
N	2048	490	263

The table presents the percentage of villages in the constituency having several public goods or facilities for the entire sample of 2,801 constituencies by reservation status. In some cases the difference between the means are significant, but we can see that the gap even if significant is not large.

Table 1.5: Quota affects Attributes of Candidates

VARIABLES	(1) All candidates	(2) Party candidates	(3) Winners	(4) Non incumbents	(5) Mean
Total Assets	-9.726***	-13.175***	-39.471***	-8.071***	16.05
	(1.846)	(1.995)	(10.064)	(1.823)	(0.44)
Immovable Assets	-7.182***	-9.145***	-29.076***	-6.060***	11.64
	(1.582)	(1.357)	(6.436)	(1.607)	(0.33)
Has a Criminal Record	-0.044***	-0.060***	-0.138***	-0.042***	0.19
	(0.012)	(0.015)	(0.048)	(0.012)	(.003)
No. of Criminal cases	-0.159***	-0.208***	-0.885**	-0.141***	0.48
	(0.048)	(0.066)	(0.371)	(0.044)	(.01)
No. of Serious Crimes	-0.129**	-0.164**	-0.678**	-0.114**	0.29
	(0.050)	(0.067)	(0.275)	(0.047)	(.01)
No. of Candidates	-2.086***	-2.103***	-1.469***	-2.109***	13.48
	(0.654)	(0.559)	(0.466)	(0.666)	(.04)
Males	-0.056***	-0.063***	-0.084**	-0.059***	0.94
	(0.009)	(0.011)	(0.037)	(0.009)	(.002)
College and Above	-0.006	-0.031	-0.058	-0.001	0.4
	(0.016)	(0.021)	(0.055)	(0.017)	(.004)
Observations	20,280	13,118	1,768	19,363	15,752

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Column 5 is the mean of the control group for all candidates. All the specifications include the percentage of Scheduled Castes population in the constituency and states as control. The estimates obtained are based on Calonico et al. (2014) which implements robust bias corrected local polynomial RD point estimators and an equal bandwidth of 6 for both sides has been considered. Candidates with total assets above the 999th quantile have been dropped to confirm the results are not being driven by candidates with exceptionally high assets. The results hold on excluding candidates in the top 1 percentile as well, but the estimates decrease slightly. The optimal CCT bandwidth was approximately equal to six for all variables. The number of observations indicate the sample within a bandwidth of six and for linear polynomial. The estimates remain similar for different polynomial specifications and selection of bandwidths.

Table 1.6: No difference in Village Facilities between SC and non-SC Constituencies

Outcome	(1) Middle School or Higher	(2) Health Centers	(3) Transport	(4) Electricity
Reserved for SC	-0.0123	0.0163	0.0083	0.032
	(0.0287)	(0.0272)	(0.0379)	(0.0348)
Observations	1490	1530	1673	1887
Reserved for SC	-0.0234	0.00087	-0.0139	-0.0133
	(0.0190)	(0.0172)	(0.0147)	(0.0191)
Observations	1345	1373	1456	1525

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

The estimates obtained are based on Calonico et al. (2014) which implements robust bias corrected local polynomial RD point estimators. Estimates with and without controlling for states are included in the bottom and top panel respectively. The estimates remain similar for different polynomial specifications and selection of bandwidths. The number of observations indicate the sample within the optimal CCT bandwidth.

Table 1.7: No difference in Village Facilities between SC and non-SC Constituencies

Outcome	(1) Credit facilities	(2) Recreation facilities	(3) Tap	(4) Phone/Post Office
Reserved for SC	-0.0140	-0.0007	0.0326	0.0071
	(0.0298)	(0.0305)	(0.0448)	(0.018)
Observations	1540	1866	1948	1840
Reserved for SC	-0.0269	0.0069	-0.0147	0.0018
	(0.0190)	(0.0189)	(0.0178)	(0.0127)
Observations	1285	1514	1522	1560

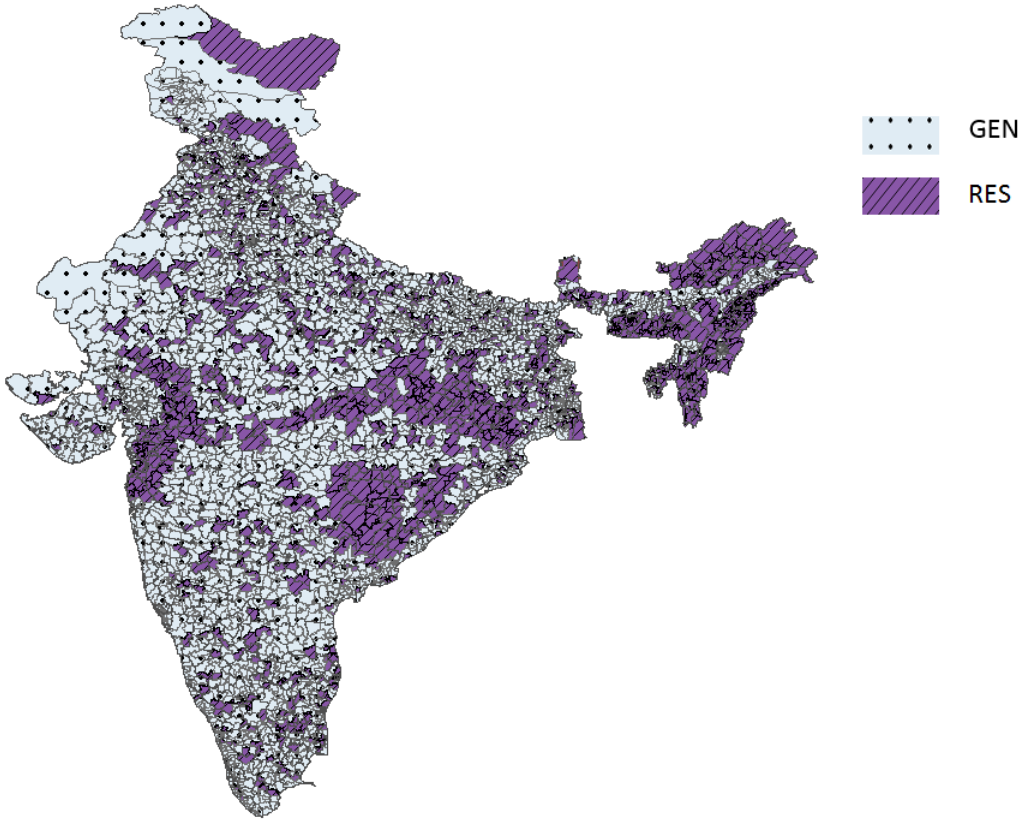
Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

The estimates obtained are based on Calonico et al. (2014) which implements robust bias corrected local polynomial RD point estimators. Estimates with and without controlling for states are included in the bottom and top panel respectively. The estimates remain similar for different polynomial specifications and selection of bandwidths. The number of observations indicate the number of constituencies within the optimal CCT bandwidth.

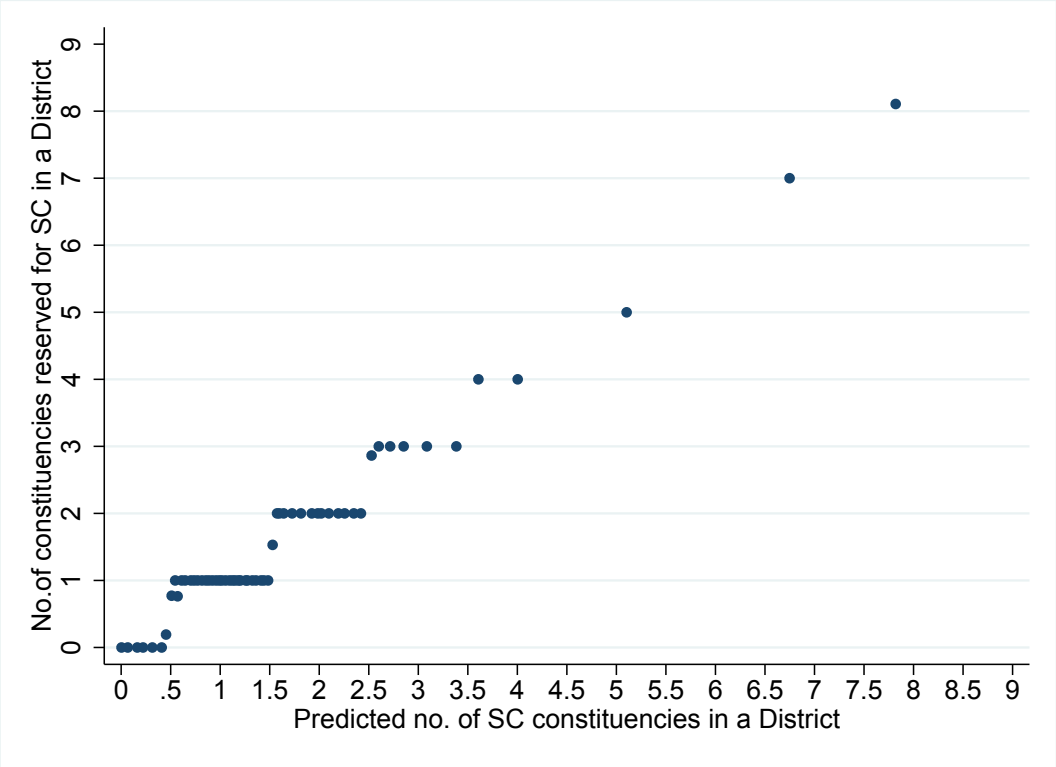
1.8 Figures

Figure 1.1: Assembly constituencies of India



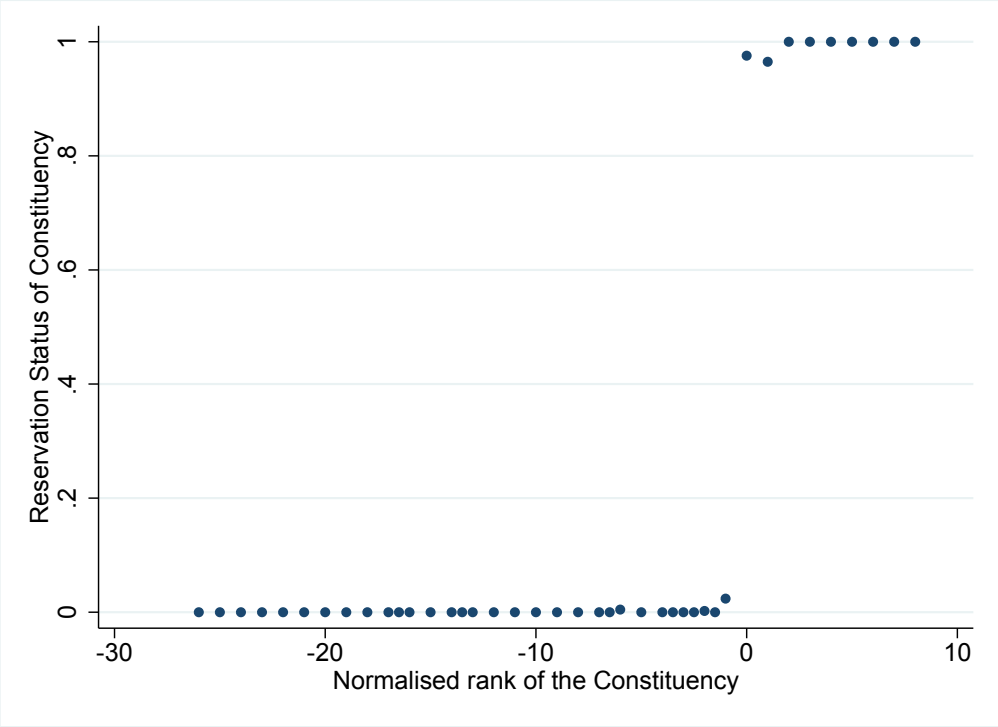
The figure represents the map of India. The dotted areas represent the unreserved or general constituencies whereas the striped ones represent the reserved constituencies.

Figure 1.2: District wise allocation of SC Constituencies



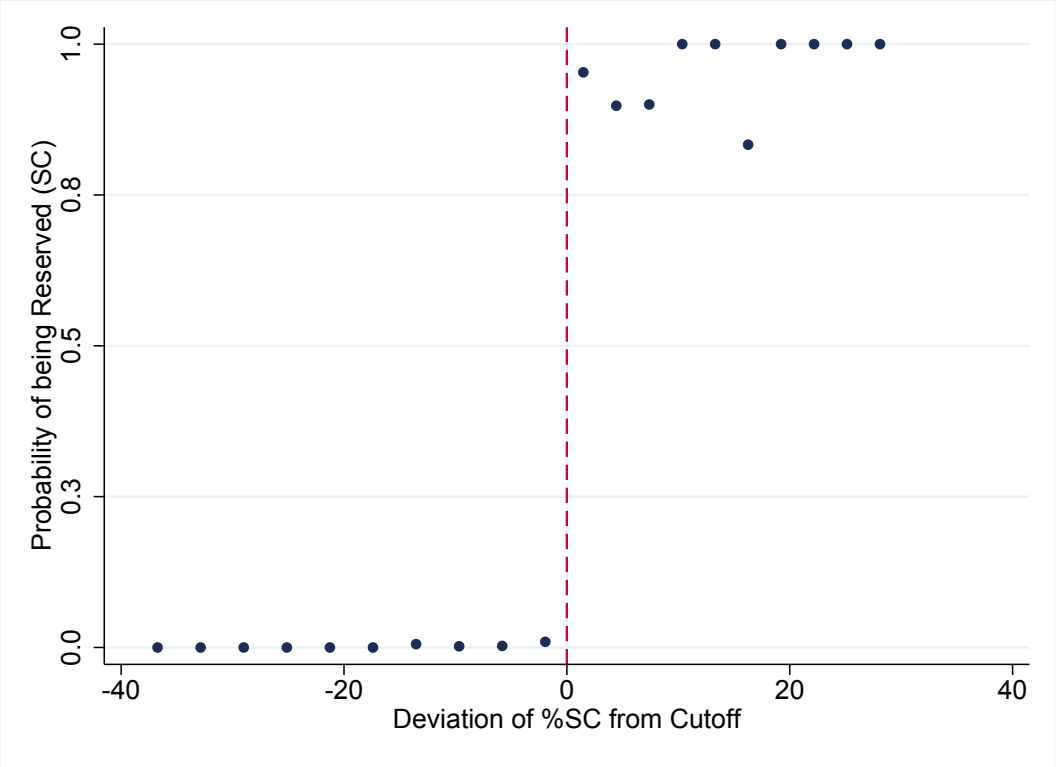
The figure plots the number of SC constituencies allocated to districts vis-a-vis the predicted number of SC constituencies that a district was supposed to receive. The figure shows that the rule followed by the Delimitation Commission for the allocation resembles a step function.

Figure 1.3: Reservation Status and Rank of the constituencies



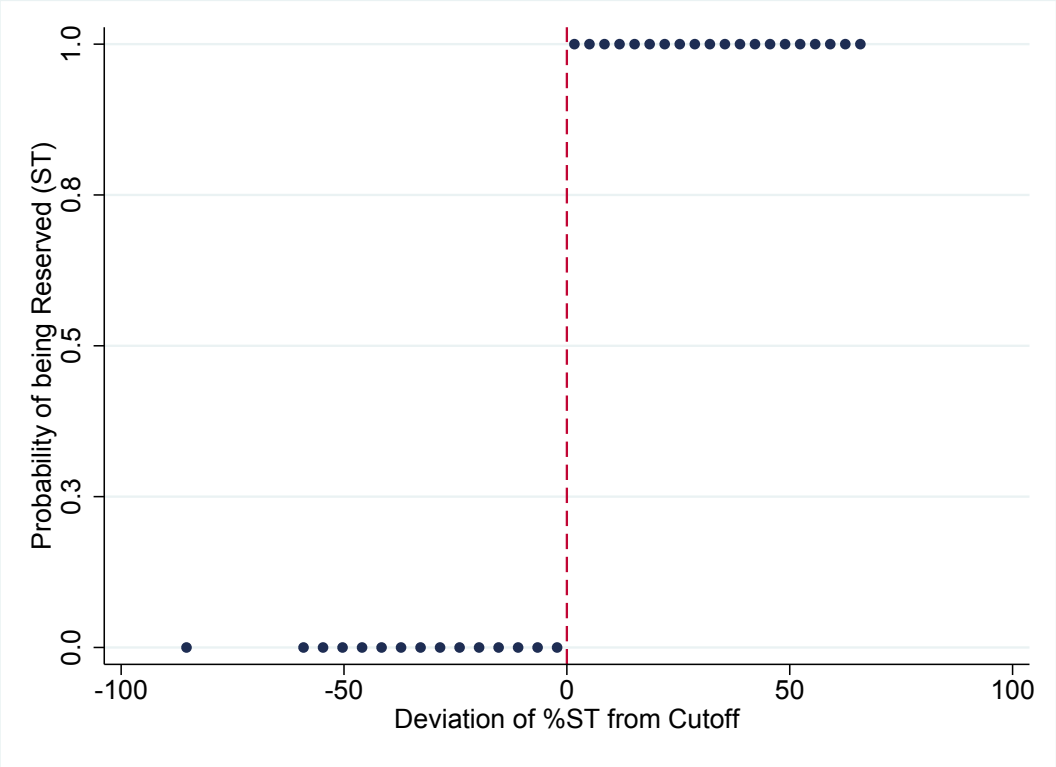
The figure plots the relationship between the reservation status of the constituency and the normalised rank of the constituency based on the percentage of Scheduled Castes population. The no. of SC seats in a district serve as the rank cutoff which has been normalised to 0.

Figure 1.4: First Stage: Scheduled Caste



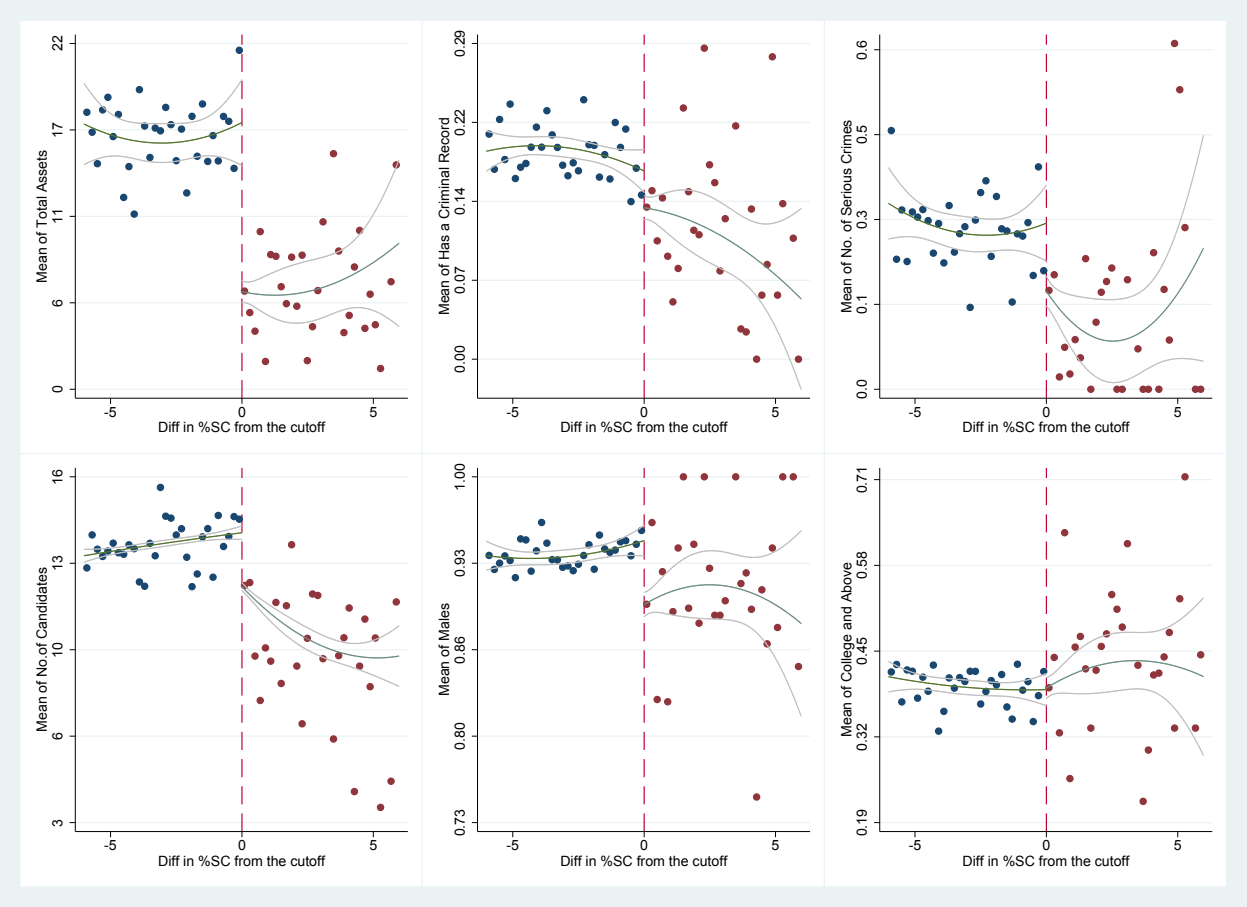
The figure plots the relation between the reservation status of a constituency for Scheduled Castes and its percentage population of Scheduled Castes. The percentage population of Scheduled Castes corresponding to the number of constituencies to be reserved acts as the cutoff for this figure. The running variable has been normalised to have the cutoff of percentage population of Scheduled Castes as 0 and thus all other points are differences of the percentage population from the cutoff. The figure shows the probability of being reserved increases by 95 percentage points on crossing the threshold.

Figure 1.5: First Stage: Scheduled Tribe



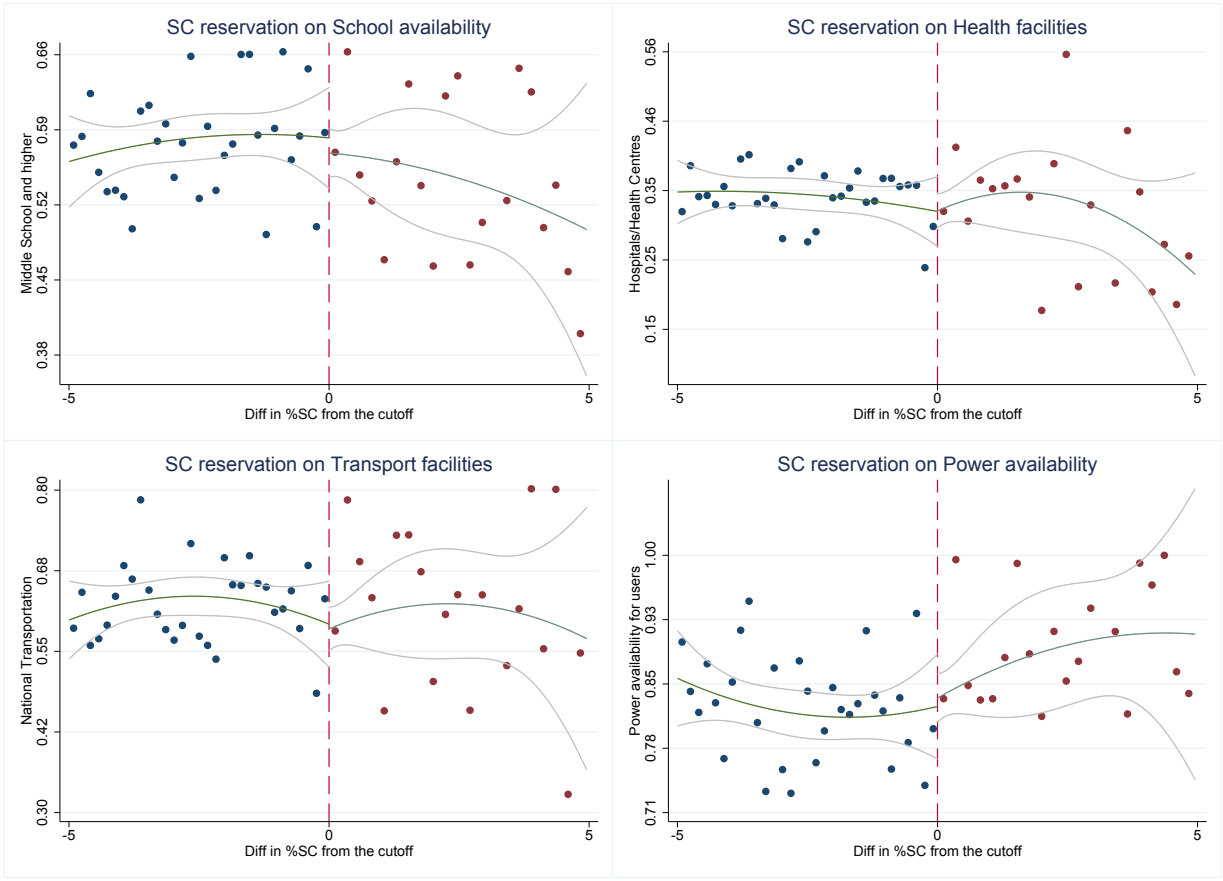
The figure plots the relation between the reservation status of a constituency for Scheduled Tribes and its percentage population of Scheduled Tribes. The percentage population of Scheduled Tribes corresponding to the rank cutoff acts as the cutoff for this figure. The running variable has been normalised to have the cutoff of percentage population of Scheduled Tribes as zero and thus all other points are differences of the percentage population from the cutoff. The figure shows the probability of being reserved is one on the right of the threshold.

Figure 1.6: Quotas affect the attributes of Candidates



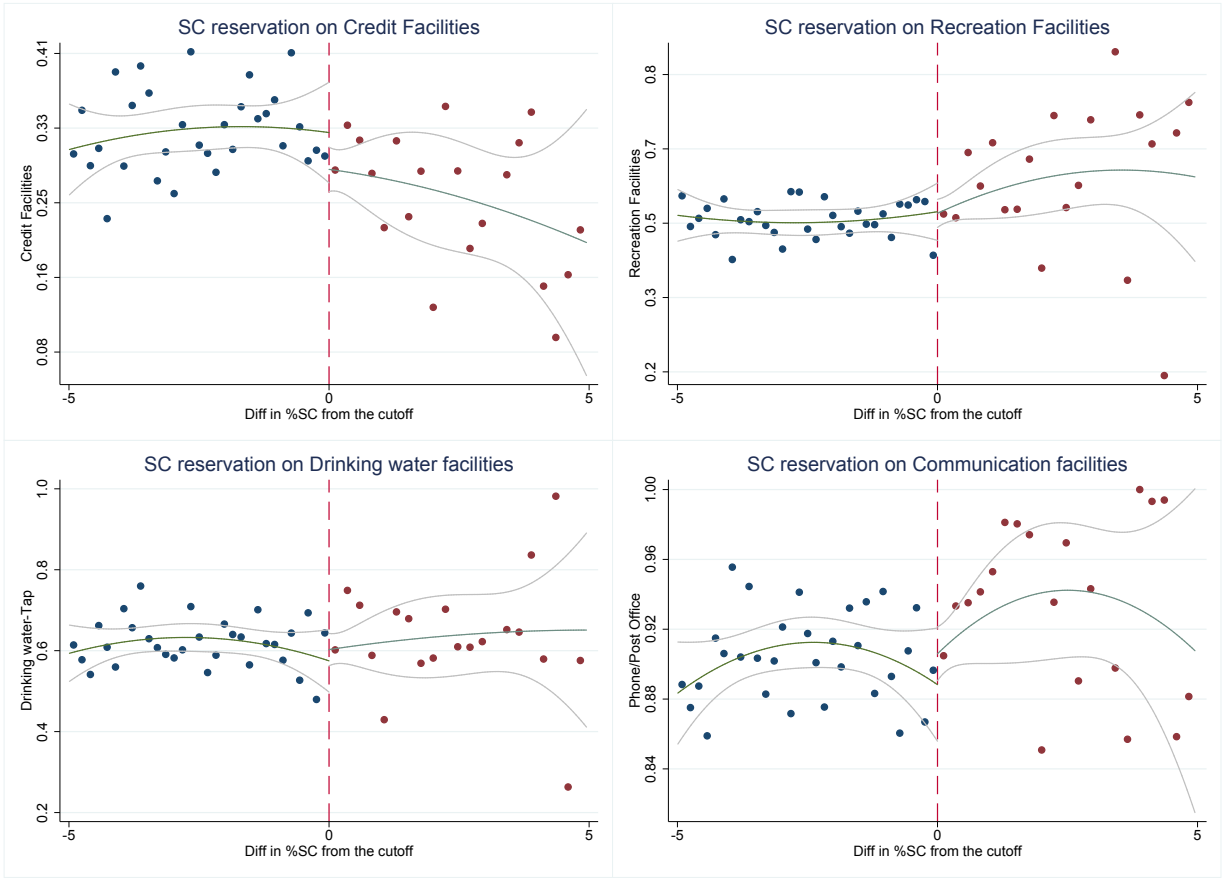
The figure plots the asset holding and criminal records of candidates based on self declared affidavits. The figure shows that the candidates contesting elections from SC constituencies have lower assets and lower criminal charges against them. There is not a significant difference in the proportion of college graduates among the political leaders on average.

Figure 1.7: No difference in Level of Village Facilities



The insignificant results remain similar on different polynomial specifications or choice of bandwidth.

Figure 1.8: No difference in Level of Village Facilities



The insignificant results remain similar on different polynomial specifications or choice of bandwidth.

Chapter 2

DOES MORE SCHOOLING INFRASTRUCTURE AFFECT LITERACY?

2.1 Introduction

Lack of schooling infrastructure has been recognized as a common problem in developing countries. Therefore, the government of these countries along with international organizations (such as the World Bank) have taken several steps to expand it. The objectives behind expansion of schooling infrastructure are to: increase accessibility of schools, promote universal primary education, or increase the literacy rate. There exists some evidence showing that investment in schooling infrastructure leads to an increase in years of education (Duflo, 2001), enrollment (Barrera-Osorio et al., 2011), and educational achievement (Case and Deaton, 1999).[1] But, increasing access through building schools can be expensive, which also requires simultaneous scaling of other infrastructure and learning resources to be effective (Muralidharan and Prakash, 2017). This becomes even more of a concern for large government programs, since other complementary resources may not be available. In this paper, I study if expansion in schooling infrastructure can influence the literacy rate, especially the literacy rate of females.

Being literate can be considered as the basic step for attaining education, and in a country with poor levels of literacy, ensuring universal literacy can be considered a priority. 750 million youth and adults in the world cannot read or write, whereas 250 million children do not have the basic literacy skills (Unesco, 2016), with 52% of the illiterate youth and adults belonging to South and West Asia (Unesco, 2016). This is also accompanied by women having even lower levels of literacy with a gender gap in literacy of 7.3% for the world (Unesco, 2016). It has also been established that a low level of literacy is associated with negative consequences, such as unemployment, poverty, and early rate of pregnancy among females (Bown, 1990; Burchfield et al., 2002;

[1]For example: An increase in educational input or pupil-teacher ratio led to an increase in enrollment and test scores (Case and Deaton, 1999). There was an average increase of 0.12 to 0.19 years of education for each primary school constructed per 1,000 children after school construction program in Indonesia (Duflo, 2001). Similarly, expansion of publicly funded private schools in Pakistan led to an increase in enrollment of 51 percentage points (Barrera-Osorio et al., 2011).

Robinson-Pant, 2006).

To study this question, I use a nationwide education program launched by the Government of India, which aimed to make primary education universally accessible and to improve the education status of girls. The ongoing program Education for All movement or Sarva Sikshya Abhiyan (SSA) entails building schools, providing necessary goods and facilities (for example: textbooks, uniforms, drinking water, and separate toilets for girls), hiring more teachers, and providing training for the teachers.[2] Additionally, the program incorporates facilities in schools and curriculum to accommodate students with disabilities and students from diverse groups of population. [3]

With a gender gap of 21.59 percentage points (Census, 2001) between male and female literacy rate in mind, the program had a special focus on girls. Focussing on improving education opportunities for girls through construction of schools (Andrabi et al., 2013; Kazianga et al., 2013) or scholarship programs (Filmer and Schady, 2008; Kremer et al., 2009) are not uncommon in developing countries. Under the SSA program there were specific schemes for girls in subdistricts or blocks called the educationally backward blocks (EBBs); EBBs were blocks with a low female literacy rate and a high gender gap in literacy rate.[4] Classification as an EBB entitled the blocks to receive additional funding to build special facilities, such as residential schools for girls, and for conducting campaigns to encourage enrollment of girls. Since, the basis of the classification was to target and improve the education facilities of areas with low female literacy rate, I study if the classification as EBB brought a significant increase in the female literacy rate.

However, the challenge in identifying the effect of being classified as an EBB on the growth in literacy rate is resolving the problem of endogeneity. For example, the blocks that are classified as educationally backward are also economically backward and have a low literacy rate originally. Thus, to find a causal estimate, we need blocks that are otherwise similar to the EBBs, but were not classified as an EBB. To solve the endogeneity issue, I use the Regression Discontinuity (RD)

[2]Understanding the effectiveness of providing a particular input has also been an important question in the literature. There have been several studies which have studied the effectiveness of providing a specific input such as school feeding program in Philippines (Jacoby, 2002), flipcharts (Glewwe et al., 2004), teachers (Banerjee et al., 2007; Duflo et al., 2012), textbooks (Glewwe et al., 2009), grants (Das et al., 2013), and cycles to schools (Muralidharan and Prakash, 2017).

[3]Prior to SSA, there have been other programs to enhance education in India whose effectiveness have also been studied in the literature like the DPEP (Jalan and Glinksya, 2013; Azam and Saing, 2017; Khanna, 2015). Mid-day meal program (Afridi, 2010, 2011; Singh et al., 2014), Operation blackboard (Chin, 2005). After 2001, SSA became the comprehensive program for many schemes before it targeting at all the different needs required for schooling.

[4]Blocks also known as Tehsils, Mandals, or subdistricts are the subdivisions of a district in India. Gender gap in literacy rates is defined as the difference between the male and female literacy rate.

method by exploiting the criteria used to classify a block as an EBB. The exercise would compare similar blocks that were barely eligible and ineligible for classification as an EBB. I use this to estimate the impact of being an EBB on the literacy rates after a decade of implementation of the program.

An important step is to investigate if the planned proposal for building schools was implemented and if there was a difference in the intensity of the program for the educationally backward blocks. To confirm whether or not there was an actual expansion in the number of schools as promised by the program, I use the school census of all schools in India. To the best of my knowledge, this is the first paper using a seven year panel dataset of around a million school in India. I find that classification of a block as an EBB led to an increase in the number of schools built in the last decade; EBBs had approximately 14 more schools (43% higher) and 0.93 more girls' schools compared to NEBBs around the cutoff. Moreover, the data shows that residential schools for girls were built only in the EBBs as specified by the program. There were improvements in various facilities, such as availability of computers, classrooms, and electricity. There has been construction of new facilities along with a repair of the previous facilties. Using data on disbursement of funds, I also find that the states with a higher proportion of EBBs received a higher funding from the program. These evidences suggest that there was a significant expansion in the schooling infrastructure and the EBBs did receive a higher allocation of resources under the program.

But, the EBBs did not show a significant change in the female literacy rate after a decade of being classified as educationally backward and receiving an expansion in schooling infrastructre. The estimate obtained is statistically zero and any effect greater than 1.2 percentage points for a sub-district can be ruled out. I find a similar insignificant result for the decrease in the gender gap in rural literacy with a confidence band of 0.7 percentage points.

The channel through which we can expect an expansion in schooling infrastructure to affect literacy is that by making education more accessible to children, having them enrolled in school would help them learn to read and write, which would increase the literacy level of the current and the future generation. But, it has also been observed that enrollment in a school may not be enough for a student to achieve any level of learning like being able to read or do math.[5] Thus, for the increase in schooling infrastructure to convert to an increase in literacy rate or educational out-

[5]For example, the 2011 ASER report for rural India states that 38.4% of students in grade one could not recognize a letter and 36.5% of students could not recognize numbers (ASER, 2011). Banerjee et al. (2010) in their experiment in Jaunpur district found similar evidence. Muralidharan (2017) suggests that improving school inputs in the "business as usual" manner will have little effect on learning outcomes, recommending changes in pedagogy and governance.

come, it would need to be accompanied by an increase in enrollment, quality teachers, and effective learning methods. A similar concern exists for other efforts which involve building infrastructure to solve a social problem, such as the Total Sanitation Campaign in India for elimination of open defecation. The campaign involved construction of toilets throught out the nation, but that did not guarantee proper functioning toilets or their use by the citizens.

It is possible that it may take more time for such an infrastructure investment to cause a change in the literacy rate, although the literacy rate in 2011 would be after a decade of the implementation of the program. Also, given the situation in India, it may take time for areas with low literacy rates to catch up, and poor transportation facilities can additionally hinder accessibility to schools. Thus, an expansion in schooling infrastructure may not lead to a quick solution if the aim is to increase literacy rate. There might be other cost effective and targeted methods for a quicker solution to low levels of literacy.

Sarva Sikshya Abhiyan is a large program that has received a lot of attention. There have been several evaluation reports and case studies about the program (PEO, 2010; Paisa, 2012), but rigorous causal empirical analysis is limited. The closest paper to this study is Meller and Litschig (2015), which finds a positive effect of two girl-specific programs in EBB called National Program for Education of Girls at Elementary Level (NPEGEL) and Kasturba Gandhi Balika Vidyalaya (KGBV) on enrollment, and also find a reduction in the gender differences in enrollment. Apart from these two programs which were supposed to be implemented only in the EBBs, I also find that EBBs received higher targeting from the SSA program, such as there were more schools built in the EBBs. Another paper (Jalan and Glinksya, 2013), however, did not find any decrease in the gender gap in enrollment or educational achievement due to the District Primary Education Program (DPEP) in India, which was an education program before the SSA. Both papers studied the effect of education programs on enrollment rates; this paper contributes to the literature by examining if similar schooling expansion programs have an effect on the literacy rate.

The rest of the paper is organized as follows: Section 2.2 describes the program. Section 2.3 outlines the empirical analysis. Section 2.4 provides the results obtained. Section 2.5 presents additional analyses, and Section 2.6 concludes the paper.

2.2 The Program

2.2.1 Aims and Objectives

Sarva Shiksha Abhiyan (SSA) is a comprehensive effort by the Government of India to tackle the various issues related to accessibility of schools in India. This is also an effort to ensure universal primary education for the nation which has been the aim of the Government of India since the nation's independence in 1947. The first formal statement on universal primary education was the National Policy on Education (1968), and regular attempts have been made to ensure universal primary education since then. As of the year 2015, 1.4 million children aged 6 to 11 were out of school in India (Benavot and UNESCO, 2015), and the dropout rate for students in grades 1 to 8 was 42.4% in 2009-10 (UNESCO, 2015). A concern for out of school children is the risk of getting involved in child labor, which is also a serious problem in many developing countries.[6] Additionally, the Government of India has taken several steps to increase the literacy and educational status of the population. The literacy rate in India has grown to 74.04% (2011 Census), but the nation still has the largest number of illiterate population in the world. Therefore, SSA was implemented in the last decade to address the various problems related to education in India.

A major hurdle in making education accessible is the absence of schools in rural areas or schools with poor infrastructure. Thus, one of the main aims of the program has been to build new schools and improve the infrastructure of existing schools.[7] Improvement in infrastructure include building and repairing classrooms, toilets, and drinking water facilities. In addition to improving the infrastructure, the program made primary education free for everyone, removing the cost of schooling as an obstacle. Finally, to improve the quality of learning and to make the curriculum more relevant, the curriculum was adapted to address children from diverse background in different locations.

Along with the low accessibility of education to the population, girls face additional challenges in enrolling or remaining enrolled in school. Some of the reasons for low enrollment or high dropout rate among girls is they are often required by their families to help with domestic chores, to take care of siblings or are married off at an early age (Glick, 2008). Other factors such as the distance to schools, lack of female teachers, the absence of girls' toilets, and safety concerns also

[6]The number of children aged between 5 to 14 involved in child labor in India is reported as 10.13 million out of 259.64 million children in the same age group (Cry.org and Census 2011)

[7]This also formed a major component of the expenditure of funds. The exact proportion directed for infrastructure or civil works varied across years and states but remained closer to 50%. For example, in the year 2006-07, 71% of the total allocation of SSA fund was towards civil works and the final expenditure was 56%. Source: State-wise and Component wise allocation and expenditure reports, SSA

contribute to lower enrollment of girls (Lloyd et al., 2005; Burde and Linden, 2013).

Thus, to encourage education for girls, the program allocates extra funds at the district level for awareness campaigns to drive enrollment, to build toilet for girls, and to train teachers to be sensitive to girls' needs. Additionally, specific schemes like the National Program for Education of Girls at Elementary Level (NPEGEL) and Kasturba Gandhi Balika Vidyalaya (KGBV) were introduced which included building residential schools for girls. These schemes are an integral component of SSA but have a separate status and receive additional funding.[8] Therefore, this program can be seen as a universal drive encompassing schooling infrastructure, but catering to gender specific needs at the same time.

To identify areas that are falling behind substantially, especially in female literacy, the Department of Education and Literacy classified blocks as educationally backward (EBB) or not educationally backward (NEBB). The classification was based on the twin criteria of rural female literacy rate being below the national average of 46.13% and the gender gap in total literacy being above the national average of 21.59%. As an exception, some blocks or urban slums with a high proportion of minorities, namely the Scheduled Castes (SCs) and Scheduled Tribes (STs), were classified as an EBB even if they did not satisfy the twin criteria.[9] The educationally backward blocks were classified to receive the girl specific schemes of NPEGEL and KGBV. Classification of blocks into EBB and NEBB forms the basis of my identification strategy.

2.2.2 Funding and Implementation of the Program

The program has a bottom-up approach in planning which makes the program more effective and relevant to the society. There are planning teams at different levels of administration, including at the district, block, and habitation level. The program collects and incorporates feedback from the local community about the problems and requirements related to schooling in the area. The program also provides responsibilities to the Village Education Committees (VECs) to monitor the schools, hire teachers, request resources from the district level administration and provide feedback to the higher level authorities.[10]

[8] KGBV was launched as a separate scheme in 2004 but has been merged into SSA from Apr, 2007. The SSA must also carry out planning for implementation of these schemes.

[9] Specifically, blocks belonging to districts with female literacy rates of the SC or ST population below 10% and comprising at least 5% of the population; Source: MHRD planning and appraisal

[10] However, there was some evidence that few parents knew of the existence and the responsibilities of VECs in the Jaunpur district of Uttar Pradesh in 2005 (Banerjee et al., 2010)

The administration of the program is decentralized with project offices being set up at local levels to help with the implementation of the program. For successful implementation, it is necessary to monitor if the tasks under the program were being carried out successfully. Hence, officials from resource centers at lower levels of administration, such as at the block and cluster levels, were to visit the schools and the sites of construction for schools. In addition, several other institutions have been involved in monitoring the different steps of the program. For example, in 2013-2015, 38 universities across different locations were appointed as monitoring institutions. The responsibility to monitor the teachers and the school administration rests with the School Management Committees under the Program and the Right to Education Act (2009). The committees consist of members representing different groups in the society, such as people from NGOs, education department, and parents. To discuss the progress of the program, and sort out any issues that the states could be facing, joint review missions are held twice a year from 2005 onwards.[11]

SSA is an ongoing program with funds being allocated by the central and state governments every year. The financing of SSA was borne by the tax base of the country with an additional levy of an educational tax. Initially, the program was to be funded entirely by the government of India, but later from 2004 onwards due to an insufficiency of funds, additional funding was provided by international aid agencies like the World Bank and European Commission. The share of funds to be released by the central and state governments was proposed to be 85:15 in 2001-02, 75:25 for the years 2002-07 and to change gradually to 50:50. The total release of funds in the decade for the year 2001-2010 has been Rs.1,25,323 crore or 27.3 billion dollars.[12] The funds from the center are disbursed to the state implementation society which is then transferred to project offices in districts.

However, there have been some criticisms of the program regarding the allocation and utilization of funds. For example, there was a lack of transparency about the procedure followed for disbursement of funds and there was an underutilization of funds allocated under the program. In spite of the fact, that funds may not have been fully utilized or were utilized with a delay, there was a significant expansion of infrastructure, which should serve as a good proxy for the expenditure

[11]The information about the program in this section has been obtained from program documents available on http://ssa.nic.in, (UNESCO, 2015; PEO, 2010), and documents on SSA framework from Ministry of Human Resource Development (MHRD, 2009). The program website has now moved to http://shagunssa.nic.in

[12]The initial release of funds for the scheme was Rs.584 crore (123.75 million dollars) in the year 2001-02 which increased more than ten times to 6,846 crore (1,510.6 billion dollars) by 2004-05 and to 30,793 crore (6.73 billion dollars) by 2010-11. The Centre:State ratio differed for some states. Audited expenditure has been Rs.120,820 crores (2.63 billion dollars). The fund under SSA was used in providing teachers salary, textbooks, civil works, management grants and other expenses

of funds, or as a proxy for work done by the program. Moreover, states with a higher percentage of EBBs seem to have received a higher allocation of funds under SSA as seen in Figure 2.1. Figure 2.1 plots the allocation of SSA funds (in million dollars) to states vis-a-vis the percentage of EBBs. The figure shows that the funding allocated and spent under SSA had a positive relation with the proportion of EBBs in a state, which suggests that the program prioritized areas with higher EBBs.[13] More details regarding fund utilization can be found in the appendix.

2.3 Empirical Methods and Analysis

2.3.1 Data and Summary Statistics

This study uses information from several datasets. The classification of blocks as an EBB/NEBB is obtained from the Ministry of Human Resource Development (MHRD). The classification was based on the female rural literacy rate and the gender gap in total literacy rate for the year 2001. For examining the growth in the literacy rate after a decade of this program, I take the population census of India for 2001 and 2011 to obtain the literacy rate for the respective years.[14] The number of blocks in 2001 was 5,463, which is taken as the base sample in the paper. Table 2.1 summarizes the information for the demographic variables from the Census of 2001 and the literacy rate in 2001 and 2011 from the respective censuses.

Since, the main aim of the program was to expand schooling infrastructure, I use the District Information System for Education (DISE) dataset to investigate whether or not there was an increase in the number of schools built and an improvement in the level of various schooling facilities. The DISE dataset is an exhaustive source of information on schools in India. For this analysis, I exclude private unaided schools and schools in urban areas. Although the dataset does contain the information on the location of schools, there does not exist a common location code to match this dataset with either the data on literacy rate for blocks from the census or the data on classification of a block from the MHRD. Hence, I had to merge the datasets using names of blocks, which can be different or can be spelled differently across datasets. Even after several rounds of matching, it was not possible to match the datasets perfectly and the final sample was reduced to 3,991 blocks

[13] The plot uses the data for the year of 2008, the relation remains similar for other years. The value of funds has been converted to US dollars based on the exchange rate for the respective year.

[14] While conducting the census, the head of the household or respondent is asked for the number of literate person in the household. In case of a doubt, the interviewer is asked to make the individuals read. The literacy rate in India is calculated by taking the percentage of literates above age 6 in the population.

from 5,463 blocks which is used for analyses in the paper.[15]

Using the data, I find that the average number of schools built in a block has increased over the years as shown in Figure B.1 of the appendix; the increase being higher in EBBs compared to NEBBs. Furthermore, such a significant difference between number of schools built in EBBs vs NEBBs is not observed for the previous decade of 1990-2000 as shown in Figure B.2. The average number of schools built annually in the rural areas of a block is around 4 to 6 schools, with some blocks having higher or lower number in certain years. The third plot shows there has been zero construction of KGBV schools in the NEBBs, which according to the program were to be built only in EBBs.[16] The results remain the same when the outcome considered is the number of schools per thousand children. To examine if there was an improvement in schooling infrastructure, I compare the level of schooling facilities such as classrooms, electrification, and computers for the academic years of 2005 and 2011 in Table 2.2. The table provides the average level of facility in a school in a block. As observed from the table, the level of inputs received by a school has increased over time. For example, the percentage of schools that are electrified has doubled, although approximately 50% of schools did not have electricity in 2011.

Additionally, we could expect the schooling condition to vary based on whether a school was old or new. For example, building the first classroom will be a priority for a new school, but an old school may focus on repairing the existing ones. Therefore, I plot the variation in the condition of schooling infrastructure by the year of establishment of the school in Figure 2.2. The different plots tell us that the availability and condition of schooling infrastructure vary across schools established in different years. We can see that as expected the proportion of good classrooms are higher in the new schools, but the new schools still need to make investments in providing electricity or computers. Thus, probably with the progress of years and increase in student capacity, the new schools will increase the number of facilities as the older schools which have been functioning for a longer period of time.

2.3.2 Identification Strategy: Fuzzy Regression Discontinuity

The challenge in estimating the effect of being classified as educationally backward on any outcome of interest is the problem of omitted variables. This is because the classification of EBBs was

[15] More details on the matching procedure is available in the data appendix.

[16] The DISE dataset has the year a school was built, and I have used the DISE dataset for 2013 to find all schools that were established until 2010. The KGBV schools are recognized based on the residential status of the school.

not random and these were blocks with a low female literacy rate. Using a simple OLS regression would lead to the following specification:

$$Outcome_i = \alpha + \beta EBB_i + \varepsilon_i \tag{2.1}$$

where EBB is a binary variable, which equals 1 if the block is an EBB and 0 otherwise. The outcome variables of interest in this paper are the literacy rate indicators and the level of schooling infrastructure of a block. However, due to the problem of omitted variables, we cannot use the method of OLS estimation as that would lead to biased coefficients.

Therefore, to tackle the endogeneity problem, I use the method of Regression Discontinuity (RD) as an empirical strategy. This is possible because the classification of EBBs was based on a criteria; rural female literacy rate below the national average of 46.13% and the gender gap in total literacy rate above the national average of 21.59% in 2001. Regression discontinuity assumes that blocks on both sides of the cutoff are otherwise similar except to exposure of the program; I provide evidences for the validity of the assumptions in the next subsection. This helps me to find an unbiased estimate of the local average treatment effect (LATE) by comparing observations only around the cutoff.[17]

The criteria for classification used two variables; rural female literacy rate and gender gap in total literacy rate based on Census 2001. This would imply two running or assignment variables. In the standard scenario, while obtaining RD estimates using local linear regressions, observations within an optimal bandwidth based on a single forcing variable is considered. Therefore, I restrict the sample to 2,626 blocks which satisfy the criteria of having the gender gap in total literacy rate (GGLR) above the national average of 21.59%; maintaining the rural female literacy rate (RFLR) for year 2001 as the single assignment variable.[18]

The relation between the proportion of blocks that were classified as an EBB and the RFLR is represented in Figure 2.3. As seen from the first panel of the figure, there does not exist a clear discontinuous relation between the probability of a block being classified as an EBB and the RFLR when all the observations are considered. But, after considering only blocks which satisfy

[17] To the best of my knowledge, the criteria used for classification of blocks as educationally backward is not used for any other classification or for implementing any other programs.

[18] The alternative of restricting the sample to ones which satisfy the RFRL threshold and using gender gap in total literacy rate as the assignment variable leads to a smaller sample. However, the results are not different as seen in Figure B.3 of the appendix.

the gender gap criteria, we observe a clear discontinuity as shown in the second panel. However, the probability of classification as an EBB does not change from 1 to 0 on crossing the cutoff of RFLR, and there are some blocks classified as an EBB even when the RFRL is above the cutoff or vice versa.

This results in a fuzzy regression discontinuity design instead of a sharp regression discontinuity. The fuzzy design occurs due to few exceptional cases, such as classification of blocks belonging to districts with at least 5% of SC or ST population and female literacy rate of the SC or ST group below 10% as an EBB, even if they did not satisfy the general criteria (MHRD, 2009). The second panel of Figure 2.3 shows the probability of a block being classified as an EBB increased by approximately 70 percentage points at the cutoff, i.e on having a RFLR less than the national average of 46.13%. The regression specification for a fuzzy regression discontinuity can be defined as follows:

$$Y_i = \alpha + \beta D_i + f(X_i) + \varepsilon_i \tag{2.2}$$

$$D_i = \alpha_1 + \beta_1 Z_i + g(X_i) + \mu_i \tag{2.3}$$

where Y_i is the rural female literacy rate or the gender gap in rural literacy rate for 2011, X_i is the rural female literacy rate for 2001, D_i=1 if a block is an EBB, and Z_i=1 if the RFLR for a block in 2001 was below the cutoff. The endogenous variable EBB is instrumented with the dummy variable Z_i, i.e the discontinuity we obtain in the first stage is used as an instrument for obtaining the causal effect of classification as an EBB.

However, estimating the above equations parametrically would rely on the choice of a functional form. An alternative method is to find estimates based on local linear regression that provides a nonparametric method of estimating the causal effect at the cutoff using local linear regression.[19] The coefficient of interest β can then be estimated as below:

$$\beta = \frac{\lim_{x \uparrow c} E[Y|X=x] - \lim_{x \downarrow c} E[Y|X=x]}{\lim_{x \uparrow c} E[D|X=x] - \lim_{x \downarrow c} E[D|X=x]} \tag{2.4}$$

by considering only observations close to the cutoff, which provides estimates that are unbiased

[19] Detailed discussion and review in Lee and Lemieux (2010).

but have lower precision. To balance the tradeoff between bias and precision of the estimates, observations within an optimal bandwidth are considered for the estimation. I report the bias corrected robust estimates using the optimal bandwidth procedure based on Calonico et al. (2014) (CCT) which measures the average treatment effect at the threshold. Additionally, I present the estimates based on Nichols (2016), which uses the Imbens and Kalyanaraman (2012) optimal bandwidth.

The first stage estimates are presented in Table 2.3, which shows that the probability of being classified as an EBB increased by approximately 75 percentage points on crossing the cutoff for the RFLR. The estimates remain consistent with different regression specifications and for different values of bandwidth.[20]

2.3.3 Regression Discontinuity Validation

The RD analysis is valid under certain assumptions: there must be no manipulation of the treatment variables around the cutoff, there is no discontinouos difference in the covariates across the cutoff, and the assignment variable is continuous. I check for the satisfaction of each of these assumptions.

Firstly, I do not find any evidence of manipulation of the treatment variable. I validate this assumption using the McCrary test from McCrary (2008), and the null hypothesis of no manipulation is accepted. From Figure B.4 we do not see any evidence of bunching of observations around the cutoff. Additionally, the RFLR and the GGLR are calculated using the census data by finding the percentage of literate population in the block. Manipulating an aggregate variable which is a ratio is not straightforward as one has to be able to manipulate the number of literate population, or the total population, or both. After the population census of India is conducted, the data is released by the Registrar General and Census Commissioner of India, which is an independent organisation, and tampering with this data is highly unlikely. Furthermore, the data on population from the census is also used for several other purposes. Thus, it is difficult to manipulate the aggregate literacy indicator of a block by anyone which could result in a misclassification.

[20]The states and Union territories of Delhi, Chandigarh, Goa, Andaman and Nicobar Islands, Sikkim, Lakshadweep, Puducherry, Daman and Diu, Dadra and Nagar Haveli had no blocks classified as educationally backward. I also perform the analysis by excluding states which did not have any block classified as EBB and the results remain similar.

Secondly, as shown in Figure B.5, covariates such as the total population of a block, the percentage of males in the population, and the population of children in the age group 0 to 6 are found to be continuous across the cutoff of RFLR for 2001. Given the rural female literacy rate and the gender gap in total literacy rates are continuous variables, the final assumption of the assignment variable being continuous is satisfied.

2.4 Results

Effect on Number of Schools

Figure 2.4 shows the relation between the number of KGBV or residential school for girls, total number of schools, schools for girls only, and coeducational schools built in the decade of 2000-2010 in a block with respect to the rural female literacy rate of the block. The figure has been constructed for observations with above average gender gap in literacy rate. First, we observe that the number of KGBV schools is zero on the right of the cutoff, which shows that the KGBVs were built only in the EBBs as directed under the program. The second plot also shows that the total number of girls' schools differed discontinuously at the cutoff. However, the number of schools exclusively for girls is much lower in general as most of the schools in the given context are coeducational schools. Such a discontinuous relation was not observed in the decade before the implementation of SSA, i.e for the years 1990-2000 as shown in Figure B.6 of appendix.[21] Second, we do observe that the total number of schools established in the last decade differed according to the eligibility status of an EBB and the trend is similar when only coeducational schools are considered.

The RD estimates obtained are reported in Table 2.4. The table shows that being classified as an EBB led to 14.45 more schools in the last decade, the coefficient is positive although insignificant with high standard errors. To underst and the relative magnitude and reduce the variance, I also obtain estimates for the logarithm of the total number of schools. The estimate obtained is 0.659 which is significant at the 5% level. Usually, there are few schools that are exclusively for girls. The coefficient for number of girls' schools and KGBV schools are positive and significant at the 1% level. Although, the coefficient implies a 200% increase compared to the mean of the control group but approximately one more school was built for girls only. It is also to be noted that EBBs on the farther left of the cutoff experienced a higher expansion in the number of schools than the EBBs closer to the cutoff. This could be due to prioritizing EBBs that have relatively lower rural

[21] According to the data, there were 15 KGBVs that were built in the last decade which is also observed in the plot.

female literacy rate, which can be a justified effort on the part of the implementer. Therefore, we observe a significant expansion of schools in the last decade, especially for the EBBs because of SSA.

Effect on Literacy Rate

But, such an expansion in the number of schools did not cause a significant change in the rural female literacy rate of the blocks. Figure 2.5 presents the RD plots for the outcome variables: increase in the rural female literacy rate and decrease in the gender gap in rural literacy rate for the last decade (2001-2011). The figure shows that there is no significant discontinuity in the outcome variables at the cutoff. Although the program did not have a significant improvement for observations around the cutoff, blocks on the left had a greater increase in the rural female literacy rate, and the effect was higher for blocks farther away from the cutoff.[22]

The point estimates for the literacy outcomes are presented in Table 2.5, which are positive but not statistically significant. The confidence interval for the estimates are relatively small implying precise estimation of the zero result. As also evident from the figure, any effect greater than approximately 1.2 percentage point increase in the rural female literacy rate over a decade can be ruled out, which would comparatively be a small effect. Similarly, any effect greater than 0.7 percentage point decrease in the gender gap in rural literacy rate can be ruled out which once again is insignificant. The insignificant results remain robust for different specifications and for different values of bandwidths.

2.5 Additional Analysis

2.5.1 Difference-in-Differences Design

To complement the RD analysis and as a robustness check, I also perform a difference-in-differences analysis. Using a difference-in-differences methodology for the entire sample requires validation of the parallel trend assumption, which unfortunately is not possible because of data limitations. However, keeping this limitation in mind, I use difference-in-differences analysis to find the average effect for all the blocks that were classified as educationally backward. The relevant equation is the following:

[22]The results remain similar for total literacy rate or the total female literacy rate as well.

$$Y_i = \alpha + \beta EBB_i + \delta Post_i + \lambda EBB_i * Post_i + \varepsilon_i \tag{2.5}$$

The year of 2001 is considered as the pre treatment period and the year 2011 is taken as the post period.

Table 2.6 provides the DID estimates for the outcome variables: the rural female literacy rate (RFLR) and the gender gap in rural literacy rate (GGRLR). The estimates imply that the educationally backward blocks experienced a 5.6 percentage points increase in RFLR in the post period as shown in column (1) on average, whereas the GGRLR declined by 0.98 percentage points, as shown by estimates in column (3). Thus, comparing all the EBBs and NEBBs, we observe there was an increase in the rural female literacy rate for the EBBs in the last decade. This could be driven by blocks farther away from the cutoff which received higher intensity of the program (had greater number of schools built as shown in Figure 2.4) and also had a higher percentage point increase in the rural female literacy rate. But, this included comparing blocks that were not similar in other characteristics as well. To validate the regression discontinuity results, I consider only observations close to the cutoff. I obtain similar insignificant result as shown in column (2) of Table 2.7. The estimate for an increase in RFLR is positive and the maximum effect would be around 1.2 percentage points which is similar to the maximum effect predicted using RD. Similarly, the estimate obtained for the GGRLR in column (4) is insignificant, but the coefficient is negative implying a decrease in the GGRLR.

2.5.2 Spillovers across Blocks

The analysis so far was to understand the effect of the program on the blocks or sub-districts, but to account for any spillover across sub-districts, I also find the effect of the program at the district level. To do this, I perform a district level analysis using the percentage of educationally backward blocks in the district. Districts are the next higher level of administration compared to blocks. The districts vary in the composition of blocks that are educationally backward, which can serve as a measure for the intensity of the program received by a district. The districts are divided into quantiles based on the percentage of EBBs. As shown in Figure B.7, there was a greater increase in the rural female literacy rate and a larger decrease in the gender gap in rural literacy rate for districts in the higher quantiles, i.e with higher percentage of EBBs.[23]

[23]I also use this measure of intensity of program for districts in chapter 3, where I explore whether the literacy rate and educational backwardness of districts are spatially correlated. I do find spatial correlation in the errors from a

I compare the result for districts obtained above with the growth in the literacy rate for the previous decade, and find that the growth was lower for districts with a higher proportion of EBBs in the previous decade. To illustrate, I plot the decrease in the gender gap in the literacy rate for 1991-2001 (before the launch of SSA) and for the period of 2001-2011 in Figure B.8. The figure shows that the decrease in the gender gap in literacy rate was lower for districts with a high proportion of EBBs in 1991-2001, implying these districts were in a worse situation compared to other districts. But, the trend is reversed in the 2001-2011 decade. There is an upward trend in the decrease in the gender gap in literacy rates, i.e districts in the upper quantiles with a higher proportion of educationally backward blocks experienced a larger decrease in the gender gap in literacy rates. Thus, educationally backward districts had a higher growth in the literacy rate during the implementation of SSA in the previous decade.[24]

2.6 Conclusion and Discussion

This paper explored the effect of expansion in schooling infrastructure and targeted education policies on the literacy rate of a nation, especially the female literacy rate. To do this, I used an education policy in India, which invested in improving schooling infrastructure with a special focus on facilitating education for girls. This study finds the program led to a huge expansion in the number of schools, an improvement in infrastructure, and facilities like residential schools for girls. Furthermore, the expansion was discontinuously higher for blocks which were educationally backward, and such an expansion was not observed in the earlier decade. But, this did not lead to a significant impact on the literacy rate after a decade of implementation of the program.

A possible concern in this analysis is that literacy rate is calculated using all population above age 6 in India, but the policy only affects (or is more likely to affect) children from age 6 to 18 (in school age). Hence, an analysis using cohorts of children during the period under study would have been more useful, which unfortunately is not possible due to unavailability of data for population belonging to different age groups in a block. However, there would be several cohorts of children from primary to secondary schools who would have benefited from the policy in the last decade. Moreover since the adult literacy rate (literacy rate of population above age 15) has also been increasing over time, if there was an increase in the literacy rate of children, it must have been

non-spatial model, which may lead to inconsistent but unbiased estimates.

[24] I was able to include all the blocks in the districts according to the respective census for the analysis in Figure B.8, since the district level analysis did not have the problem of incomplete matching of blocks across several datasets.

captured as an increase in the overall literacy rate.

Nevertheless, investment in schooling infrastructure along with teaching and learning methods should form a strong foundation for improving education and literacy in the long run for future generations. When implemented in the form of large scale programs, which has been common in India and other countries as well, special efforts need to be made regarding scaling other resources and ensuring precise implementation of the program. Alternatively, there have been other local based efforts to increase adult literacy, such as Sahajani Shiksha Kendra by a national NGO Nirantar, litreacy classes by Mission India, a missionary organisation. Comparing the effectiveness of such local based programs might be an interesting area for future research.

2.7 Tables

Table 2.1: Summary Statistics for the sample

	EBB		NEBB	
Variable	Mean	Std. Dev	Mean	Std. Dev.
%Scheduled Caste	16.27	8.25	17.84	10.81
%Males	51.196	1.28	51.01	1.42
%0 to 6 child	17.84	3.08	14.34	2.48
%Scheduled tribe	15.02	23.78	11.21	20.934
Rural Female Literacy Rate'01	33.43	8.04	57.94	9.23
Rural Female Literacy Rate'11	48.34	7.37	67.16	8.827
Gender Gap in Total Literacy Rate'01	28.1	5.7	20.24	6.73
Gender Gap in Total Literacy Rate'11	22	4.8	15.2	5.76
No. of blocks	1,770			2,221

EBB: Educationally Backward Block, NEBB: Not Educationally Backward Block. The above table summarizes some of the key demographic variables from Census 2001.

Table 2.2: Growth in School Inputs

	Year 2005	Year 2011
No of visits by officials from CRC	5.38	5.90
	(0.06)	(0.07)
No of visits by officials from BRC	2.00	2.46
	(0.03)	(0.03)
Amount of SSA fund Received (Rupees)	4018.69	13523.37
	(85.00)	(204.13)
Amount of SSA fund spent (Rupees)	3518.15	12134.43
	(72.69)	(101.84)
No of classrooms in a school	3.34	3.86
	(0.02)	(0.02)
%Schools having electricity	25.11	51.22
	(0.44)	(0.60)
No of computers in a school	0.35	1.09
	(0.02)	(0.03)
No. of blocks	3935	3973

The table summarizes some of the variables for schooling infrastructure using the DISE data for the two academic years of 2005 and 2011. The variables represent the average input for a school in a block. CRC: Cluster Resource Center, BRC: Block Resource Center.

Table 2.3: First Stage Estimates: Probability of being classified as EBB

Variable	(1)	(2)	(3)
Rural Female Literacy Rate	-0.754***	-0.727***	
	(0.0759)	(0.0838)	
Rural Female Literacy Rate			-0.761***
			(0.0555)
Observations	327	327	776
Control	N	Y	N
Bandwidths	IK	IK	CCT

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Column (2) provides the estimates obtained on controlling for demographic variables and state fixed effects. The controls used are percentage of Scheduled caste and Scheduled tribe population, percentage of males, and percentage of children in the age group 0 to 6 in rural area based on Census 2001 data. I report estimates using both the optimal bandwidth criteria of Imbens and Kalyanaraman (2012) and Calonico et al. (2014) which are 2.6 and 6.1 respectively. The IK estimates were obtained using the code of Nichols (2016) and the CCT estimates were obtained using the code of Calonico et al. (2014). The estimates obtained remain similar for different values of bandwidth or polynomial specification.

Table 2.4: Effect of being classified as EBB on number of schools built

VARIABLES	Schools	Girls' schools	KGBV schools	Coed Schools
EBB	14.45	0.932***	0.239***	13.53
	(9.324)	(0.299)	(0.0730)	(9.172)
Mean of Control	32.580	0.463	0.018	32.192
	(1.800)	(0.075)	(0.007)	(1.787)
Observations	843	681	958	846

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The number of observations are within the optimal Calonico et al. (2014) bandwidth. Standard errors have been clustered at the district level.

Table 2.5: Effect of being classified as EBB on literacy rates

VARIABLES	(1) IRFLR	(2) IRFLR	(3) IRFLR
EBB	0.406	0.110	
	(1.010)	(0.605)	
EBB			0.430
			(1.163)
Mean of Control			10.742
			(.25)
Observations	743	743	856
	DGGRLR	DGGRLR	DGGRLR
EBB	0.823	0.253	
	(0.570)	(0.352)	
EBB			0.778
			(0.568)
Mean of Control			6.35
			(.124)
Observations	618	618	891
Bandwidth	IK	IK	CCT
Control	N	Y	N

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The top panel represents estimates in percentage points for increase in rural female literacy rate (IRFLR) of a block from 2001 to 2011. The estimates for the second outcome variable of interest, decrease in gender gap in rural literacy rate (DGGRLR) which is defined as Male rural literacy rate-Female rural literacy is provided in the bottom panel. DGGRLR is constructed by subtracting the gender gap in rural literacy rate of 2011 from gender gap in rural literacy rate of 2001. The column specifications are similar to the specifications in Table 3.

Table 2.6: DID estimates for Literacy rates

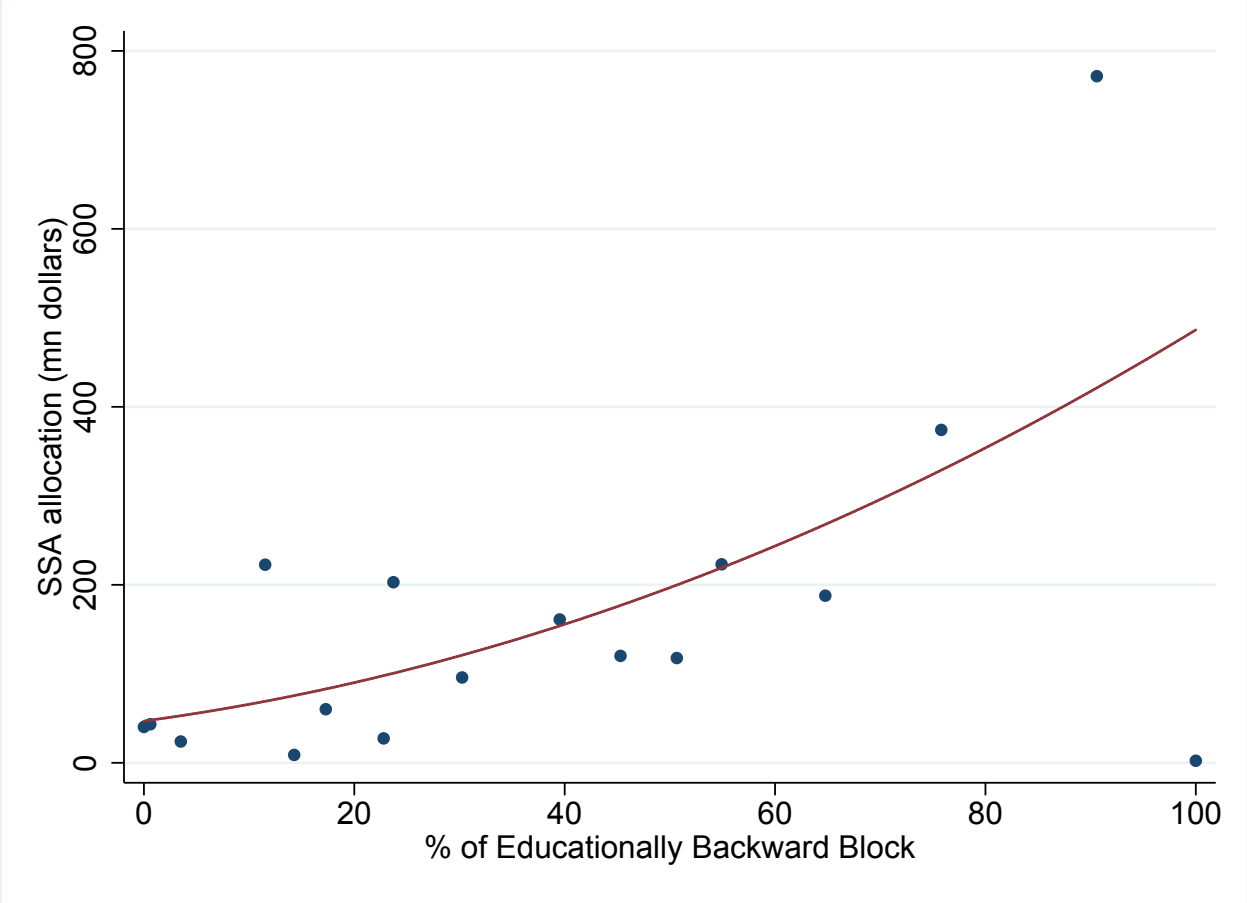
VARIABLES	(1) RFLR	(2) RFLR	(3) GGRLR	(4) GGRLR
EBB*Post	5.660***	0.920	-0.983**	-0.277
	(1.224)	(0.589)	(0.417)	(0.300)
R-squared	0.832	0.848	0.764	0.870
No. of blocks	3,845	742	3,845	742
Block Fixed effect	Y	Y	Y	Y

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

RFLR: Rural Female Literacy rate, GGRLR: Gender gap in Rural Literacy rate. Column (1) and column (3) provide DID estimates for the whole sample, whereas, the estimates in columns (2) and (4) are obtained for the sample of blocks close to the cutoff. The estimate is significant for the entire sample, but not for the sample close to the cutoff. However, the coefficients are positive and negative implying an increase in the Rural Female Literacy rate and decrease in the Gender gap in Rural Literacy rate respectively. The regressions include block fixed effects and standard errors are clustered at the state level.

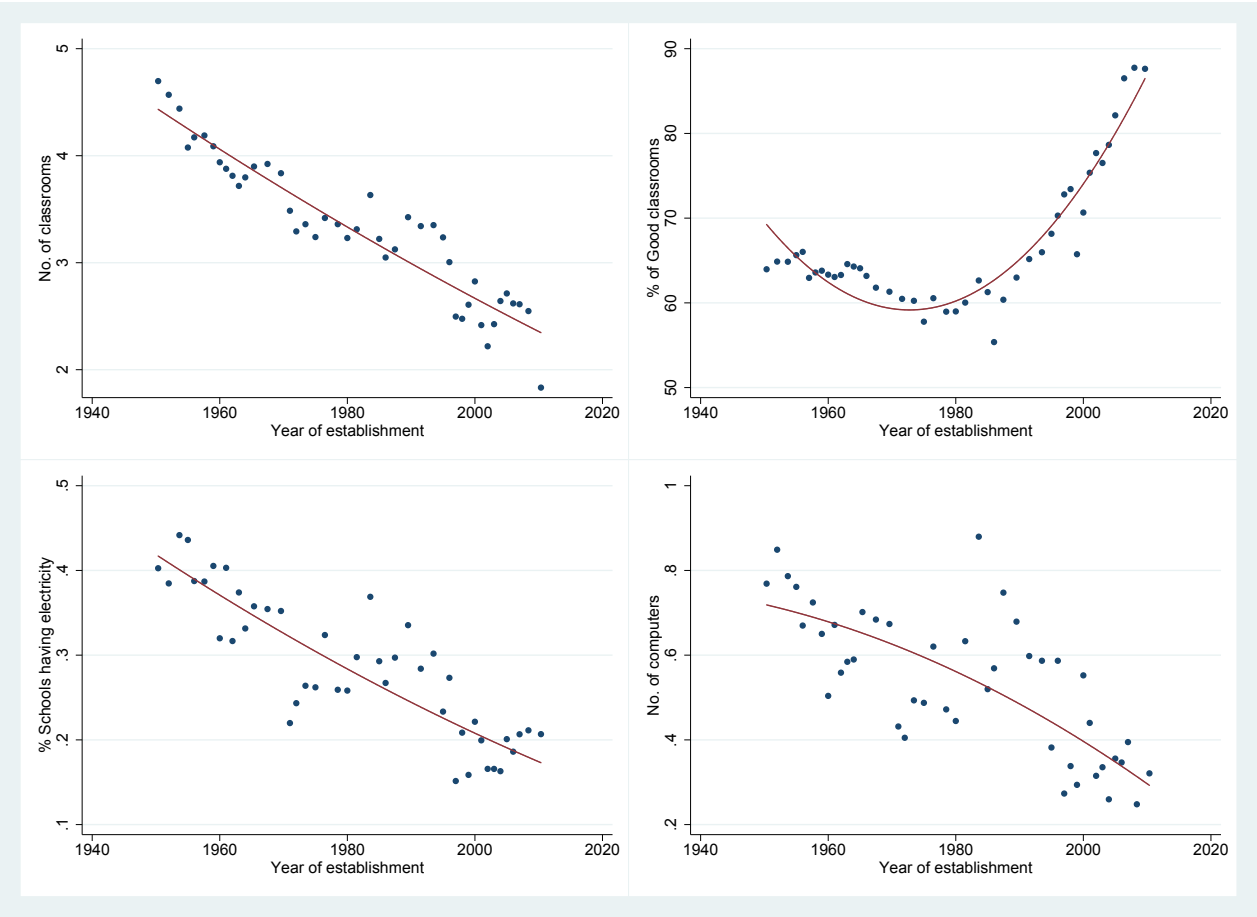
2.8 Figures

Figure 2.1: Amount of SSA funds allocated to states



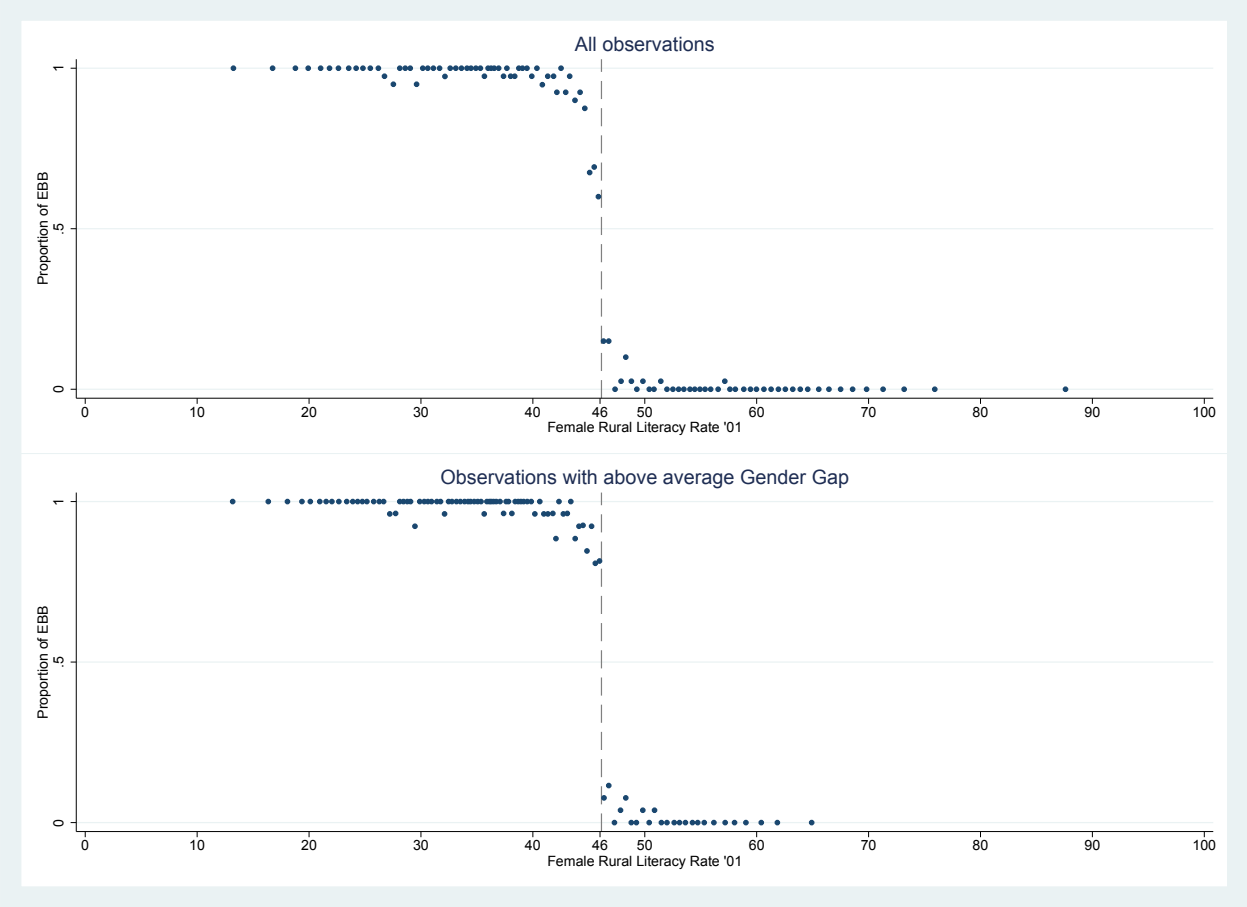
The figure shows a positive relation between the quantity of SSA fund allocated to a state and the percentage of Educationally Backward Blocks in the state

Figure 2.2: Situation of Infrastructure by year of establishment of school



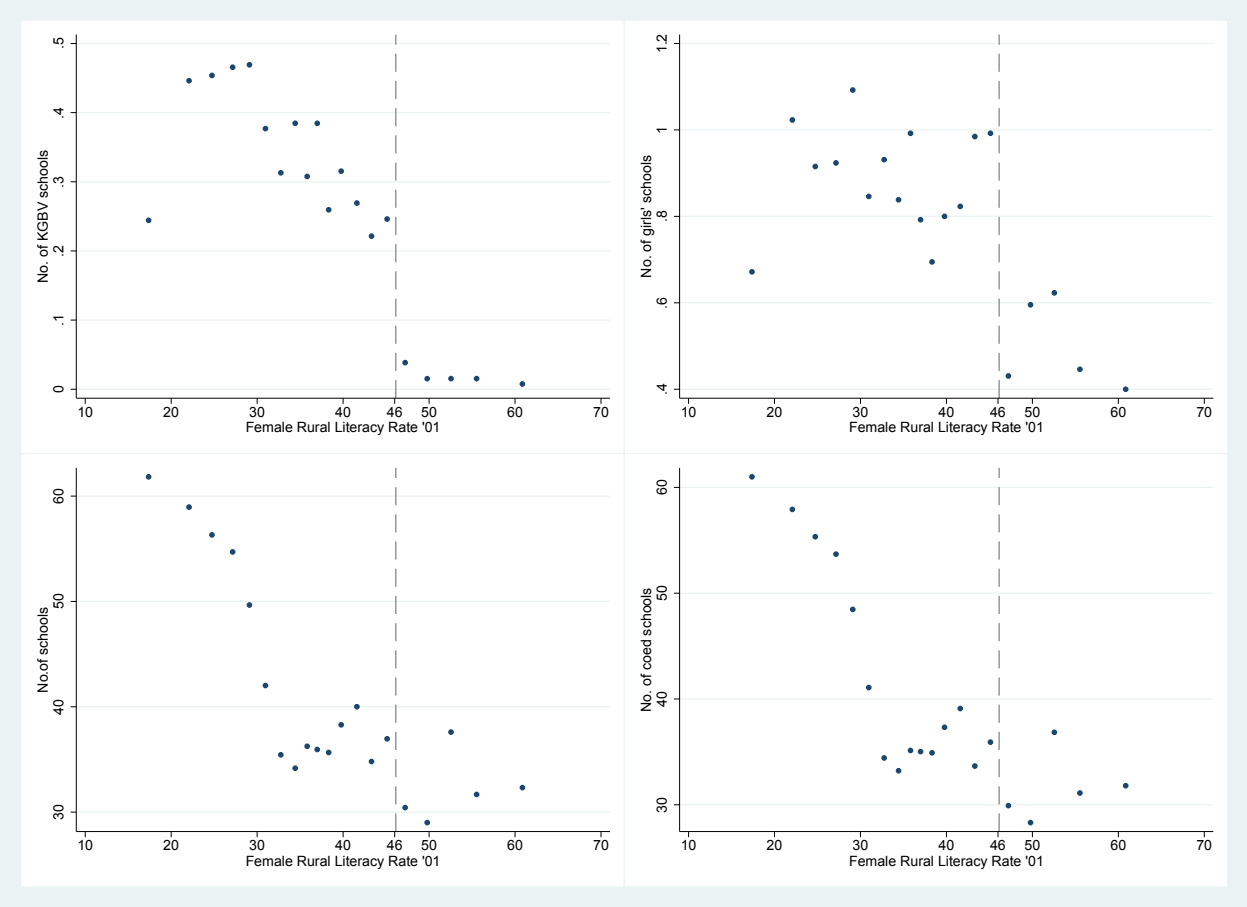
The figure shows that the proportion of good classrooms are higher in the newer schools but probably will need more time to catch up with the old schools in the number of computers or electricity. This could also be because the old schools have been functioning for longer period of time and have higher enrollment and capacity.

Figure 2.3: Regression Discontinuity for first stage



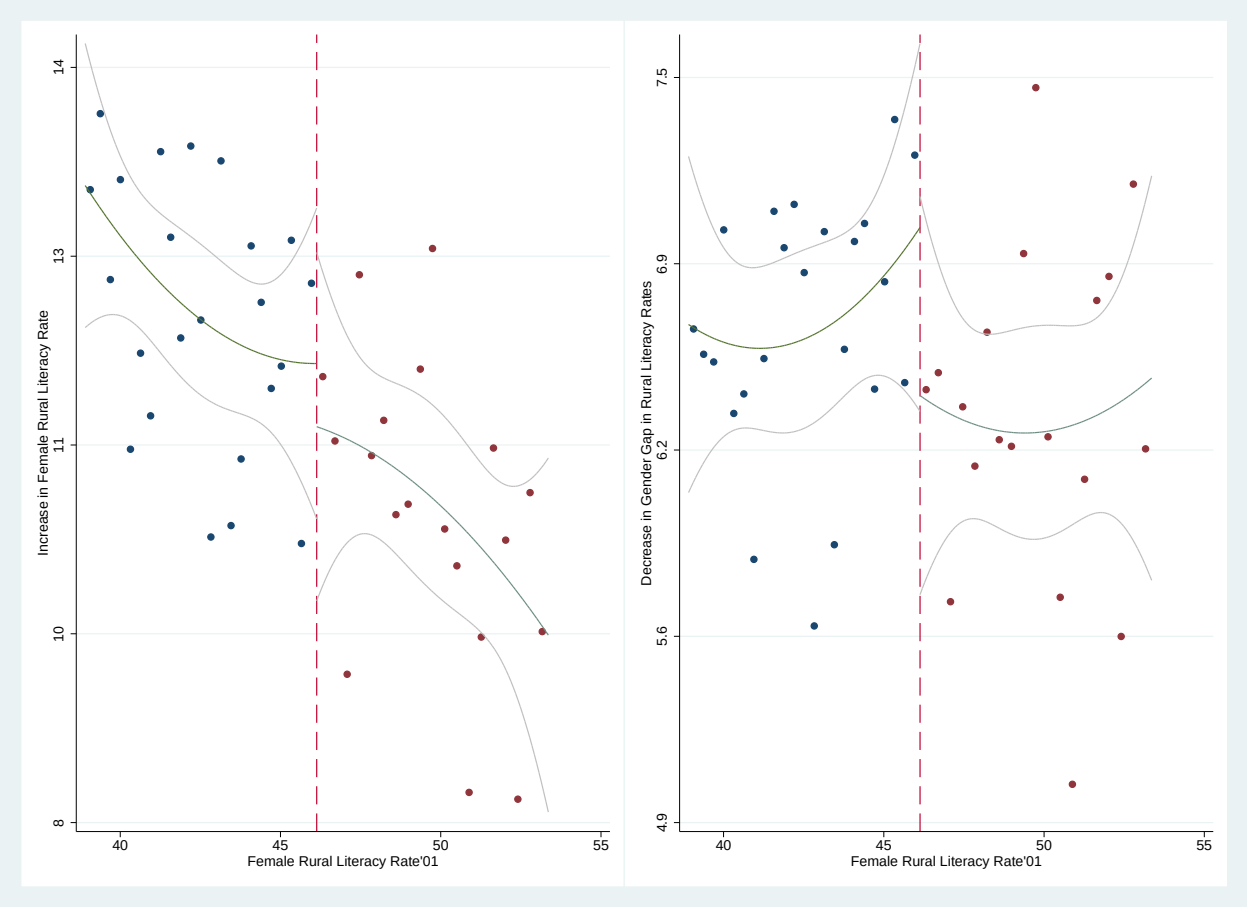
The figure shows that the probability of being classified as EBB increased by 70 percentage points at the cutoff. The plot on the top is for the entire sample whereas the plot on the bottom is obtained by resticting observations that satisfy the criteria of having the gender gap in total literacy rate above the national average.

Figure 2.4: Schools built in the last decade



The number of schools established in the last decade (2001-2011) differed discontinuously according to the eligibility status of EBB and the number of KGBV schools established drops to 0 on crossing the cutoff.

Figure 2.5: Effect on literacy indicators



The figure plots the increase in rural female literacy rates and the decrease in gender gap in rural literacy rates over the decade for the optimal bandwidth calculated based on Calonico et al. (2014). The graphs show that there was no significant discontinuity at the cutoff.

Chapter 3

SPATIAL ANALYSIS OF AN EDUCATION PROGRAM AND LITERACY IN INDIA

3.1 Introduction

A common method used for the implementation of public programs to address access to basic necessities, such as education and health, is to target geographic areas in need of the public program. It is often of interest to measure the effect of such programs. Given the spatial contiguity of geographic areas, there is possibility of spatial correlation between areas that receive the program and in the outcome variables used to measure the results of the program. In standard program evaluation, the possibility of spatial dependency among geographic neighbors is usually not taken into account. Using a non-spatial model when a spatial model is appropriate would violate the classical assumptions required for obtaining biased and inconsistent estimates. In this paper, I investigate the presence of spatial correlation, and use estimates from a spatial and a non-spatial model to explore the association between expansion of schooling infrastructure and literacy in India.

I study the education program Sarva Sikshya Abhiyan (SSA) or the Education for all movement, which was launched in India in 2001. The program was a nation-wide initiative for building schools and providing other necessities to students, such as textbooks, uniforms, drinking water, and hiring new teachers. Given the low literacy level of females in India (according to Census 2001, the average female rural literacy rate in 2001 was 46.13% and the average gender gap in literacy rate was 21.59%), the program was implemented with a focus on girls. The SSA incorporated specific schemes for girls, such as building residential schools and enrollment campaigns to encourage admission of girls in schools. The program focused on districts with a high gender gap in literacy rates and a low rural female literacy rate, that is, districts that were relatively educationally backward. A more educationally backward district received higher funding from SSA, and also received additional funding for specific schemes, such as building residential schools for girls and funds for enrollment campaigns.[1] Hence, I study if there was an increase in the rural

[1]The program focused on educationally backward blocks and I am using the percentage of educationally backward blocks in a district as the measure of the intensity of the program as explained in the paper. The percentage of EBBs or intensity ranges from zero to hundred. I am also referring to districts with higher percentage of educationally backward

female literacy rate and a decrease in the gender gap in literacy rates in districts that were more educationally backward.

Spatial correlation is said to exist if the variable in one district is dependent on the variable of a nearby district. Such dependency may exist because of geographic, demographic, administrative, or any reason which can be related to distance. I examine whether or not there is spatial auto-correlation in the educational backwardness of districts, which determine the intensity of the program received, and if the outcome variables of interest (rural female literacy rate) are spatially dependent. This is in contrast to the scenario when a variable or feature may be distributed randomly in space. Spatial correlation unlike temporal correlation is not unidirectional and thus defining the direction of the spatial correlation and which units are considered as "nearby" units (units that are spatially correlated) is not obvious. To do this, I perform exploratory spatial data analysis and also find the optimal spatial weight matrix to define "neighbors", that is, the districts who are spatially dependent. I use geo-spatial data on districts of India for spatial analysis and modeling.

I find a positive spatial correlation between districts that are educationally backward, as indicated by the Moran's I statistic for global spatial correlation. Additionally, using local indicator of spatial autocorrelation (LISA), I find spatial clusters, which identify regions of spatial concentration of districts that are educationally backward, in the north-east and the center of southern India. These regions have generally experienced lower rate of development in India. There is spatial correlation in the dependent variable, the change in rural female literacy rate as well. The spatial maps also suggest evidence of a positive relation between districts that experienced a greater increase in the rural female literacy rate and districts that were educationally backward or received the program with higher intensity.

The data suggests spatial auto-correlation in the errors, therefore, estimating a non-spatial model would lead to statistically inconsistent estimates of the association between the percentage of educationally backward blocks and literacy rate. On estimating an OLS model to predict the influence of educational backwardness of a district on the literacy rate, I find the errors are not normal, are heteroscedastic, and are spatially correlated. To account for the spatial dependency, I use the Lagrange multiplier test (Anselin, 1988) to choose between the two major categories of spatial models, the spatial lag and the spatial error model, and I find the spatial error model as the appropriate model category. The spatial error model incorporates spatial correlation in the errors. I also use the likelihood ratio test to determine the spatial model which best explains the data.

blocks as a district that is relatively more educationally backward.

The resulting optimal model is the spatially distributed error model (SDEM), followed closely by the spatial error model, and I provide estimates from both the SDEM and the spatial error model. The SDEM includes spatial dependency among neighbors in the errors as well as the independent variable, which is the educational backwardness of a district or a measure of the intensity of the program received (treatment). Thus, I can also investigate whether or not there is an additional influence on the literacy rate of a district apart from the treatment it receives just because its neighbors receive a different intensity of treatment.

The SDEM estimates imply a 0.076 percentage point increase in the rural female literacy rate with a one point increase in the the educational backwardness of a district, and a 0.02 percentage point decrease in the gender gap in rural literacy rate with a one point increase in the educational backwardness of a district. The SDEM estimates were similar to the estimates from the spatial error model. The difference between the SDEM and OLS estimates are very small for the outcome increase in rural female literacy rate (0.076 compared to 0.073) and for the outcome decrease in rural gender gap in literacy rates (0.02 compared to 0.0175). The coefficient capturing spatial dependency in the educational backwardness of a district, that is the association between the treatment received by the neighboring districts and the outcome of a district is 0.011 and is insignificant. Thus, I did not find evidence of the treatment received by the neighboring districts influencing the literacy rate of a district, and the source of spatial correlation was in the errors (as also suggested by the error model).

Spatial dependency and interactions among neighbors has been an important question in the literature and is observed in various contexts, such as in education, government expenditure, environmental policy, politics, and demography. In the context of schooling, there is evidence on: spending by schools in a district to depend on school spending of neighboring districts (Ajilore, 2011; Ghosh, 2010), positive correlation in adopting open enrollment policy between neighboring school districts (Brasington et al., 2016), teacher salaries to be highly influenced by salaries in similar districts (Greenbaum, 2002), geographical contiguity to foster local competition which increases efficiency for public and private schools (Millimet and Collier, 2008; Misra et al., 2012; Gonzalez Canche, 2014; McMillen et al., 2007).

Strategic interaction has also been observed for property-tax competition among local governments (Brueckner and Saavedra, 2001), expenditure of state governments (Case et al., 1993), spending decisions of local governments in Portugal (Costa et al., 2015), recreational and cultural services provided in municipalities in Sweden (Lundberg, 2006), in incumbent behavior of

politicians (Besley and Case, 1992).[2] Similar questions of dependency has been studied for stringency of environmental policies (Fredriksson and Millimet, 2002b), pollution abatement expenditure (Fredriksson and Millimet, 2002a), and sex ratio of a district (Echávarri and Ezcurra, 2010).

3.2 Context and Data

3.2.1 Program and Measure of Treatment

The Sarva Sikshya Abhiyan or the Education for All program was launched in 2001 to increase access to education in India. This was an effort in the direction to achieve universal elementary education, which has been a goal of the government since India's independence in 1947. The program aimed to improve the schooling infrastructure in the country. Improvement in infrastructure and schooling facilities included building and repairing classrooms, building girls' toilets, and drinking water facilities. The program also hired new teachers and trained them, and also designed the curriculum to include the interests of children from diverse backgrounds. In addition to providing the necessary infrastructure, the program aimed to increase enrollment and reduce drop out rates of children.

The financing of SSA was borne by the tax base of the country with the implementation of an educational tax. The total allocation of funds in 2001-2010 was Rs.1,25,323 crore or 27.345 billion dollars and the audited expenditure was Rs.120,820 crores (2.63 billion dollars). The fund from the central government was transferred to the state, which was then transferred to the districts.[3] The program followed a bottom-up approach in planning and sought more community involvement with planning teams at the district, block and habitation level to accommodate location specific issues and needs. These teams include people from NGOs, teachers, education department, parents among others.

Given the low enrollment or high dropout rate of girls, the program made special efforts to increase the enrollment of girls in schools in India.[4] India has a large number of illiterate people and a low level of female literacy (average female literacy rate was 46.13% according to Census

[2]For an overview of strategic interaction among governments refer to (Brueckner, 2003).

[3]The funding at the central level followed a rule but the exact rule used to allocate the funds to the next administrative levels remain to be investigated.

[4]The dropout rate for adolescent girls in India is as high as 63.5% (Information from Cry.org which is based on MosPI,2012).

2001). To identify areas that are falling behind substantially in female literacy, the Department of Education and Literacy of India classified blocks (also known as sub-districts, which are at a lower administrative level compared to districts) as educationally backward (EBB) or not educationally backward (NEBB). The classification was based on the twin criteria of rural female literacy rate being below the national average of 46.13% and the gender gap in total literacy being above the national average of 21.59%. Classification as an EBB entitled the blocks to receive additional funding to build special facilities, such as residential schools for girls known as Kastruba Gandhi Balika Vidyalay (KGBV), and for conducting campaigns to encourage enrollment of girls under the National Program for Education of Girls at Elementary Level (NPEGEL).

The program was targeted to blocks, and I take advantage of the method of classification of educationally backward blocks to investigate the effect of expansion in schooling infrastructure on literacy using the method of regression discontinuity in chapter 2. I find a significant expansion in the schooling infrastructure in the educationally backward blocks, but I do not find a significant effect of being classified as EBB on the increase in rural female literacy rate. Using box plots I do find suggestive evidence of districts with higher percentage of EBBs having an increase in the rural female literacy rate and a decrease in the gender gap in literacy rate.[5]

However, in this paper I am interested to explore spatial correlation in dependent and independent variables, which may exist due to demographic, geographic, administrative reasons, or any reason related to distance, and I use the data on the literacy rate and the percentage of educationally backward in districts from chapter 2 to study this. To understand the spatial correlation, I consider districts as the relevant spatial unit; districts are a higher administrative level compared to blocks, and on average there are ten blocks in a district. I define the percentage of educationally backward blocks in a district as the treatment intensity of the program in that district.

There can be various reasons to expect spatial dependency in educational backwardness and literacy rates between districts. Firstly, areas that are educationally backward may overlap with areas that are also economically backward and geographic regions which have not experienced growth and development. Secondly, districts in the same state may benefit from common state programs or the allocation of funds and resources of a district may depend on the resources received by nearby districts. Finally, children in a district may attend schools in neighboring districts which have better schools or infrastructure, or also be motivated to enroll in school if the children in nearby districts are enrolling. This is not uncommon in rural India where distance to schools can be a major deciding factor in choosing the school to attend. There also exists evidence for insti-

[5]For further details on the analysis, please refer to (Jogani, 2018).

tutions and schooling districts engaging in strategic spending and competing for students in other countries (Ajilore, 2011; Ghosh, 2010; Greenbaum, 2002).

3.2.2 Data

I use the spatial maps of the administrative districts of India from Census 2001 shared by the Datameet community. I obtain information on the classification of blocks as educationally backward or non-educationally backward from the Ministry of Human Resource Development (MHRD). The classification was based on the rural female literacy rate and the gender gap in total literacy rate for the year 2001 based on the population Census. For examining the growth in the literacy rate after a decade of this program, I use the population census of India for 2001 and 2011. The total number of districts used for the analysis is 576 out of a total number of districts of 592.[6] Table 3.1 provides the summary statistics of the treatment variable (percentage of EBBs) and the literacy variables for districts in India in 2001. The table shows that the mean percentage of EBBs in a district is 45.9. Figure 3.1 shows the frequency distribution of EBBs and I find the median percentage of EBBs in a district to be 42.9.

3.3 Spatial Econometric Methodology

3.3.1 Detecting Spatial Auto-correlation

To investigate the presence of spatial correlation, I perform some exploratory spatial data analyses using quantile maps. For the analysis in this paper, I use the GeoDa software, which has wide applications in the field of regional science for spatial analysis (Anselin, 2005; Anselin et al., 2006). There are several analytical methods to understand local and global spatial correlation. One of the common statistic used to understand global and local spatial correlation is the Moran's I statistic. The global Moran's I statistic inform us about the overall spatial auto-correlation in a data set. However, due to spatial heterogeneity, the degree of correlation may vary across space, and thus local statistics help to determine the spatial correlation in different areas across space. The local Moran's I statistic provides the degree of association for each spatial unit and its spatial

[6]The total number of districts in 2001 was 592. However, there was no unique identification code for matching the data-sets, and thus I was able to match 576 districts accurately. The total number of sub-districts or blocks was 5,463 according to Census 2001.

neighbors. I use global and local Moran's I to understand the global and local spatial correlation in the data. A positive auto-correlation implies the spatial adjacent areas have similar values, whereas a negative auto-correlation implies the spatial adjacent areas have different values. The Global Moran's I can be represented by the following equation:

$$I = \frac{N\Sigma_i\Sigma_j W_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\Sigma_i\Sigma_j W_{ij}(x_i - \bar{x})^2} \tag{3.1}$$

where N is the total number of observations (districts in this case), i and j represent spatial units (or districts), X is the variable of interest for detecting spatial correlation (for example the percentage of educationally backward blocks or change in literacy rate), $\overline{X}$ is the mean of X, W_{ij} is the optimal weight matrix (which is discussed in the next section).

The statistic can be interpreted as a ratio of the covariance between spatial units and the total observed variance. I ranges from -1 to 1, where -1 implies perfect negative spatial correlation, 1 implies perfect positive spatial correlation, and 0 implies no spatial correlation. To determine whether or not there is significant spatial correlation is to compare the statistic I obtained from the above equation with the expected value of I under the null hypothesis of no spatial correlation, which = $-1/(N-1)$ (Dall'erba, 2005). If I is greater (lower) than $-1/(N-1)$, then the data suggests positive (negative) spatial correlation. In this analysis, $N = 576$, therefore, $-1/(N-1)$=-0.00174

The local Moran's I statistic for each region i can be obtained using the following equation (Anselin, 1995):

$$I_i = \frac{N(x_i - \bar{x})\Sigma_j W_{ij}(x_j - \bar{x})}{\Sigma_i(x_i - \bar{x})^2} \tag{3.2}$$

where I_i is the local Moran's statistic for spatial unit i, and j represents the neighboring units, N is the total number of observations (districts in this case), X is the variable of interest for detecting spatial correlation (for example the percentage of educationally backward blocks or change in literacy rate), $\overline{X}$ is the mean of X, W_{ij} is the optimal weight matrix (which is discussed in the next section).

To represent local spatial correlation, I use the local indicator of spatial auto-correlation (LISA) cluster map. The LISA cluster map presents areas with significant and insignificant local Moran's I statistic. Thus, the cluster map also helps to identify local spatial clusters or hotspots.

3.3.2 Weight Matrix

Spatial correlation implies units i and j are correlated, i.e $Cov(X_i, X_j \neq 0)$. However, with N number of observations, the number of correlations to estimate would be $N(N-1)/2$, which can be a very large number. The number of correlations to be estimated can also be reduced once we know the neighbors that interact with each other. To do this, I define a spatial weight matrix which imposes a structure on the nature of correlation between the spatial units. A spatial weight matrix usually relies on the distance between neighbors and is thus exogenous. I describe the spatial weight matrices I use below:

Distance Based Spatial weights: One of the common method is to use the great circle distance (or arc distance, which is calculated using the latitudes and longitudes of the spatial units), where $W_{ij} = 1$ if the distance between i and j is below a user defined threshold (Dall'erba, 2005; Anselin, 2005). For example, spatial units i and j are defined as neighbors if the distance between them is below the threshold of 250 miles. However, the distance has to be above the minimum distance required for every spatial unit to have at least one neighbor. I find the minimum distance required for every spatial unit to have at least one neighbor for the data set as 270 miles. I find distance based spatial weight matrices for the minimum distance (270 miles). I also find distance based spatial weight matrices for 300 miles and 325 miles to check for robustness, the results remain similar on using the different distances.

Higher order contiguity: The units i and j are said to be contiguous of the order K if the maximum number of borders to cross to reach j from i is K (Anselin, 2005). The contiguous relations can be of various kinds which lead to different weight matrices, such as matrices Queen, Rook, and Bishop. For example, in the below table, we can define the following weight matrices based on different method of selection of neighbors:

X	Y	X
Y	Z	Y
X	Y	X

Rook: The spatial units labeled as Y are considered neighbors.
Queen: The spatial units labeled as X and Y are considered neighbors.
Bishop: The spatial units labeled as X are considered neighbors.

K-Nearest Neighbors: The value of K is chosen a-priori which is used to define the K nearest

neighbors of j, the definition is based on the distance between the centroids of the spatial units. $W_{ij} = 1$ if the centroid of area i is one of the K nearest neighbor from j, $W_{ij} = 0$ otherwise (Anselin, 2005).

I use the above weight matrices to find the weight matrix which captures the nature of spatial correlation best for the given data set, or has the highest value for the local Moran's I statistic.

3.3.3 Determining the Spatial Model

I am interested in investigating the presence of spatial dependency and in estimating the relationship between the educational backwardness of a district and the increase in rural female literacy rate (or the decrease in gender gap in total literacy rates).[7] A non-spatial model can be represented by the following equation:

$$Y_i = \alpha + \beta T_i + \gamma X_i + \varepsilon_i \tag{3.3}$$

where T_i is the percentage of educationally backward blocks (EBBs) in a district, (X_i) are other demographic characteristics of the districts which are used as controls, and Y_i is the outcome of interest. In the context of this paper the outcomes are the literacy rate variables, increase in rural female literacy rate (IRFLR) and the decrease in rural gender gap in literacy rates (DGGRLR). The coefficient of interest is β.

The most common form of spatial models are the spatial lag model and the spatial error model. In case of the spatial lag model, the outcome in a district depends not only on the independent variables or the covariates of that district, but also on the outcome of its neighbors, whereas in the spatial error model, the error term of the district is correlated with that of the neighbors.[8] Using matrix notation, the spatial lag model and the spatial error model can be represented by equations (3.4) and (3.5) below:

$$Y = \rho WY + X\beta + \varepsilon \tag{3.4}$$

$$Y = X\beta + \varepsilon, \varepsilon = \lambda W\varepsilon + \mu \tag{3.5}$$

where $\mu \backsim N(0, \sigma^2 I)$.

[7] For review of spatial models refer to (Anselin, 2002).

[8] The models in the spatial literature can also be compared with those in the time series literature, such as the spatial lag model can be compared with the AR(1) model which is represented as $Y_t = \rho Y_{t-1} + \varepsilon_t$.

However, in presence of spatial correlation, a non-spatial model may lead to biased and inconsistent estimates depending on the type of spatial dependency. To illustrate, consider the spatial lag model as shown by equation (3.4). (3.4) can be rewritten as below:

$$Y = (I - \rho W)^{-1} X\beta + (I - \rho W)^{-1}\varepsilon \tag{3.6}$$

$$\implies \frac{\partial Y}{\partial X} = (I - \rho W)^{-1}\beta \neq \beta$$

Hence the estimates are biased.

Similarly, in case of a spatial error model, the errors are correlated, violating the assumption for consistency of OLS estimates.

Thus, we must define spatial models that would capture spatial dependency in the data to obtain unbiased and consistent estimators.

To choose between the spatial lag and error model, I use the Robust Lagrange Multiplier test (Anselin, 1988, 2005; Anselin et al., 1996).

Some other common spatial models are the spatial lag of X (SLX) model and the spatial durbin error model (SDEM) (Vega and Elhorst, 2015). These models are a variant of the spatial lag and spatial error models and are represented by the following equations:

$$Y = \rho WX + X\beta + \varepsilon \tag{3.7}$$

$$Y = \rho WX + X\beta + \varepsilon, \varepsilon = \lambda W\varepsilon + \mu \tag{3.8}$$

Compared to the spatial lag model, which included spatial dependency in Y as shown by equation (3.4), the SLX model includes spatial dependency in X. In the SLX model, the outcome Y of a district depends on the covariates of the district and also the covariates of the neighbors. The SDEM model is a combination of the SLX model and the spatial error model, it includes spatial dependency in X and spatial dependency in the error term. I will also be using the above models to compare the results.

3.4 Results

3.4.1 Spatial Correlation: Exploratory Spatial Data Analysis

To understand the spatial distribution of the educationally backward areas, I draw a quantile map which divides the districts of India into four groups based on the concentration of educationally backward blocks in the district. This is presented in Figure 3.2. Districts with the darkest shade are ones with 90-100% educationally backward blocks. The quantile map shows districts with high intensity of treatment were surrounded by districts with high intensity of treatment as well, and vice versa. Thus, this indicates presence of spatial correlation among neighboring districts in the proportion of educationally backward areas or the intensity of treatment.

Figure 3.3 presents the quantile map of the outcome variable of interest, increase in rural female literacy rate (IRFLR), for the period of 2001-2011. The districts with the darkest shade have experienced a maximum increase in the rural female literacy rate in the last decade. There exists a similar pattern of spatial correlation between neighboring districts, that is, districts with high IRFLR are surrounded by districts with high IRFLR and vice versa. Additionally, comparing Figures 3.2 and 3.3, we observe a correlation between districts with a high percentage of EBBs and districts with a high IRFLR. But, there are also some districts, for example in the west of India, which have a high percentage of EBBs, but a low IRFLR.

To define the spatial neighbors, I find the weight matrix that would best capture the spatial autocorrelation in the data, that is, choose the matrix with the highest Moran's I. I calculate the Moran's I (which is a measure of spatial correlation) for different kind of matrices as described in Section 3.2, that is matrices defined on the distance based spatial weight, on contiguity, and on K-nearest neighbors. On trying the different spatial weight matrices, I find the average Moran's I is highest for the K5 weight matrix for both the independent and dependent variables (EBB and IRFLR). Thus, I use the K5 matrix as the weight matrix to define the spatial neighbors for the rest of my analysis.

Figure 3.4 presents the univariate global Moran's I plot for the treatment variable, percentage of educationally backward blocks. The figure captures the spatial correlation in the treatment variable between a district and it's neighbors (represented by the spatially lagged variable on the vertical axis). The Moran's I obtained for the relation is 0.66, which is greater than -0.00174 (the value for $-1/(N-1)$ as described in section 3.1). Thus, this indicates a significant positive spatial autocorrelation between the neighbors. Figure 3.5 presents the univariate global Moran's I scatter plot

for the outcome variable IRFLR and the Moran's I statistic obtained is 0.51, which also indicates a positive spatial correlation in IRFLR between the neighboring districts.

The positive Moran's I and the linear relation shows that there is a significant correlation among spatial units and their neighbors with respect to both the outcome (increase in rural female literacy rate) and the independent variable (percentage of EBBs). Additionally, the null hypothesis of such correlation to be random, is rejected at a significance level of 1% for both variables. Figures 3.3 and 3.4 suggest the presence of global spatial auto-correlation.

To understand the nature of the local spatial auto-correlation, I present the local indicator of spatial auto-correlation (LISA) cluster map for the treatment variable in Figure 3.6 and for the outcome variable in Figure 3.6. The districts with high (low) EBB are labeled as high (low) in Figure 3.6 and districts with high (low) IRFLR are labeled as high (low) in Figure 3.7. The plots are constructed using standardized values on the axes, such that each unit corresponds to a standard deviation. The scatter plot is centered on the mean to divide the plot into four quadrants. The top right and bottom left correspond to positive spatial autocorrelation, whereas the bottom right and top left correspond to negative spatial autocorrelation.

The LISA cluster map shows the spatial clusters or hotspots, which identifies regions where there is higher spatial auto-correlation between districts. The red regions are the high-high regions, that is, districts with high values of the treatment or outcome variable are also surrounded by districts with high values of the treatment or outcome variable implying positive local spatial correlation. The blue regions are the low-low regions, which shows areas with low values of the treatment or outcome variable are surrounded by districts with low values of the treatment or outcome variable, also implying positive local spatial correlation. A High-Low region implies negative spatial correlation as High districts are surrounded by Low districts. The other regions do not show a significant presence of local spatial auto-correlation between the districts at a significance level of less than 5%.

3.4.2 Model Estimation

I begin by estimating the OLS model as shown in equation (3.3), where the independent variable is the percentage of blocks that are educationally backward in a district. I also include other covariates as controls, such as the percentage of minority population in the district during the implementation of the project (areas with high minority population have been historically disad-

vantaged and experience lower literacy rates) and also the percentage of female population. I find the Moran's I statistic of the residuals from the OLS model for the outcomes IRFLR and DGGRLR. The statistic obtained was 0.42 and 0.47 for IRFLR and DGGRLR. The Moran's I statistic was significant at a p-value of .001 with 999 permutations, implying presence of spatial auto-correlation in the residuals. Hence, to account for the spatial correlation, I use different spatial models as described in section 3.3, and incorporate the any association of the neighbors using K5 weight matrix.

Table 3.1 presents the statistics from the Lagrange multiplier test for the spatial error and lag models (as represented by equations (3.4) and (3.5)). The Lagrange multiplier statistics for the error and lag models are significant but the statistic from the Robust Lagrange multiplier test is significant only for the error model. Thus, the optimal spatial model belongs to the category of spatial error model as suggested by the Robust Lagrange multiplier test.

Additionally, I also test for the SLX model which implies dependency in the independent or treatment variable as shown by equation (3.7). But, the residuals remain spatially correlated, and therefore, I estimate the SDEM which incorporates correlation among errors along with the dependency in the treatment variable as shown by equation (3.8). Table 3.2 provides the log likelihood results, which compares the performance of the different spatial models. The Likelihood Ratio test however does not imply a significant difference in the spatial error model and SDEM for either of the outcomes, but they provide a better fit than the OLS model as indicated by a lower absolute value of the ratio.

Table 3.3 presents the estimates from both the models suggested by the Lagrange Multiplier test, the SDEM and the spatial error model, and also compares the results with the estimates obtained from the OLS model. The table shows that the estimates imply a 0.076 percentage point increase in the rural female literacy rate with a one percentage point increase in the intensity of treatment. However, the coefficient estimate for WX (or the Weighted % of EBBs), which captures the influence of the treatment received by the neighbors is insignificant. Thus, the results do not suggest a significant association between the intensity of the treatment or the program received by the neighboring districts and the literacy rate of district.

Table 3.4 provides the estimates from the various models for the second dependent variable, decrease in rural gender gap in literacy rates (DGGRLR). The estimates imply a 0.02 percentage point decrease in the gender gap in rural literacy rates with a one point increase in the intensity of treatment. The coefficient estimate for WX (or the Weighted % of EBBs) is insignificant.

Finally, to examine if the association vary between districts based on their intensity of treatment, I divide my sample into two groups: low and high intensity. I divide the districts based on the median value of the intensity of treatment. Districts for which the treatment variable, that is, the percentage of blocks that are educationally backward in a district is lower (higher) than the median belong to the group low (high). This exercise will help to capture heterogeneity in the association.

Table 3.5 presents the SDEM estimates for the low and high districts separately. The table shows that the influence of educational backwardness of a district is larger for districts in the high group compared to districts in the low group, that is the districts which were more educationally backward and hence received the program with higher intensity. However, the influence of the educational backwardness of the neighboring districts, which is measured by the Weighted % of EBBs term remains insignificant even for the districts in the high group, implying there is no association between the treatment received by the neighbors.

3.5 Conclusion

Several nations and states direct public resources to geographic areas that are lacking the basic amenities for development. There is a high possibility of spatial correlation and dependency between neighboring areas in the independent or dependent variables. Not accounting for spatial dependency may lead to biased and inconsistent estimates while studying the impact of the program. Hence, it is important to investigate if there is presence of spatial correlation and if the dependency is of a type that may bias the estimates.

In this paper, I use an education program in India which involved building schools and improving the schooling infrastructure and explore spatial dependency in the treatment and the outcome variables of interest. The program targeted geographical districts that were educationally backward. The educationally backward districts had a low level of rural female literacy rate and a high level of the gender gap in literacy rate. I also estimate the association between the program and literacy using a spatial and a non-spatial model to compare if the spatial dependency in the variables affect the measurement of the association.

I used the Moran's I statistic to measure the global spatial correlation and the LISA cluster maps to detect local spatial correlation. The data suggests evidences for presence of both global and local spatial correlation. The LISA maps also suggest spatial clusters of both high and low regions,

especially in the north-east and west of India, which are historically characterized by low levels of development and literacy. To account for the spatial correlation, I use the Lagrange multiplier and the likelihood ratio test to determine the appropriate spatial model and I use the spatially distributed error model (SDEM) as the model of choice. I also obtain estimates from the OLS model without accounting for any spatial dependency.

The SDEM estimates are similar to the OLS estimates and there is no additional influence due to the neighboring districts. In this case, the source of spatial dependency was the spatial correlation in the errors, and the value of the estimate remains similar. It is recommended to be wary of spatial dependency and use a spatial model in addition to incorporating any other controls or sources of endogeneity.

3.6 Tables

Table 3.1: Summary Statistics for Districts in India in 2001

Variable	Mean
% of EBBs	45.954
	(1.683)
Rural Female Literacy Rate	47.417
	(0.635)
Gender Gap in Literacy Rate	24.158
	(0.324)
No. of Districts	581

% of EBBs: Percentage of Educationally Backward Blocks in a district. The table presents the summary statistics using the Census 2001.

Table 3.2: Lagrange Multiplier test for Spatial Model Selection

Statistic	Δ Rural Female Literacy Rate	Δ Gender Gap in Literacy
LMError	290.8 ***	351.4***
LMLag	247.6***	328.86***
RLMError	44.3***	22.9**
RLMLag	1.2	0.35
SARMA	291.9***	351.8***

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Δ Rural Female Literacy Rate: Increase in rural female literacy rate, Δ Gender Gap in Literacy: Decrease in rural gender gap in literacy rates, LM: Lagrange multiplier, RLM: Robust Lagrange multiplier. The table presents the statistics from the Lagrange multiplier test for the spatial error and lag models. The Lagrange multiplier statistics for the error and lag models are significant but the statistic from the Robust Lagrange multiplier test is significant only for the error model. Thus, the suggested model belongs to the category of spatial error model.

Table 3.3: Likelihood ratio test to determine the optimal spatial model

Outcome	SDEM	ERROR	OLS
Δ Rural Female Literacy Rate	-1552.3	-1553.8	-1660.9
Δ Gender Gap in Literacy	-1181.9	-1184.4	-1308.9

SDEM: Spatial Durbin Error Model, ERROR: Spatial Error model, OLS: Ordinary least squares, Δ Rural Female Literacy Rate: Increase in rural female literacy rate, Δ Gender Gap in Literacy: Decrease in rural gender gap in literacy rates. The likelihood ratio in the above table is highest for the SDEM, closely followed by the Error model. Thus, I choose SDEM as the optimal model.

Table 3.4: Estimation Results for Increase in rural female literacy rates

	OLS	SDEM	Error
Variables	Estimate	Estimate	Estimate
% of EBBs	0.073***	0.076***	0.074***
	(0.004)	(0.006)	(0.005)
Weighted % of EBBs		-0.011	
		(0.01)	
No. of Districts	576	576	576

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

SDEM: Spatial Durbin Error Model, ERROR: Spatial Error model, OLS: Ordinary least squares, % of EBBs: Percentage of Educationally Backward Blocks in a district, Weighted % of EBBs: Weighted matrix of Percentage of Educationally Backward Blocks in the neighboring districts. The above table compares the estimates from the different models. The SDEM estimates are slightly greater than the OLS estimates (0.076 compared to 0.074). The SDEM estimate implies a 0.076 percentage point increase in the rural female literacy rate for a one point increase in the intensity of treatment or the educational backwardness of a district.

Table 3.5: Estimation Results for Decrease in rural gender gap in literacy rates

	OLS	SDEM	Error
Variables	Estimate	Estimate	Estimate
% of EBBs	0.02***	0.0175***	0.017***
	(0.002)	(.003)	(0.003)
Weighted % of EBBs		-0.003	
		(0.006)	
No. of Districts	576	576	576

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

SDEM: Spatial Durbin Error Model, Error: Spatial Error model, OLS: Ordinary least squares, % of EBBs: Percentage of Educationally Backward Blocks in a district, Weighted % of EBBs: Weighted matrix of Percentage of Educationally Backward Blocks in the neighboring districts. The above table compares the estimates from the different models. The SDEM estimates are similar to the OLS estimates for decrease in rural gender gap in literacy rates and it implies a 0.02 percentage point decrease in the rural gender gap in literacy rates for a one point increase in the intensity of treatment or the educational backwardness of a district.

Table 3.6: Effect by different spatial regimes

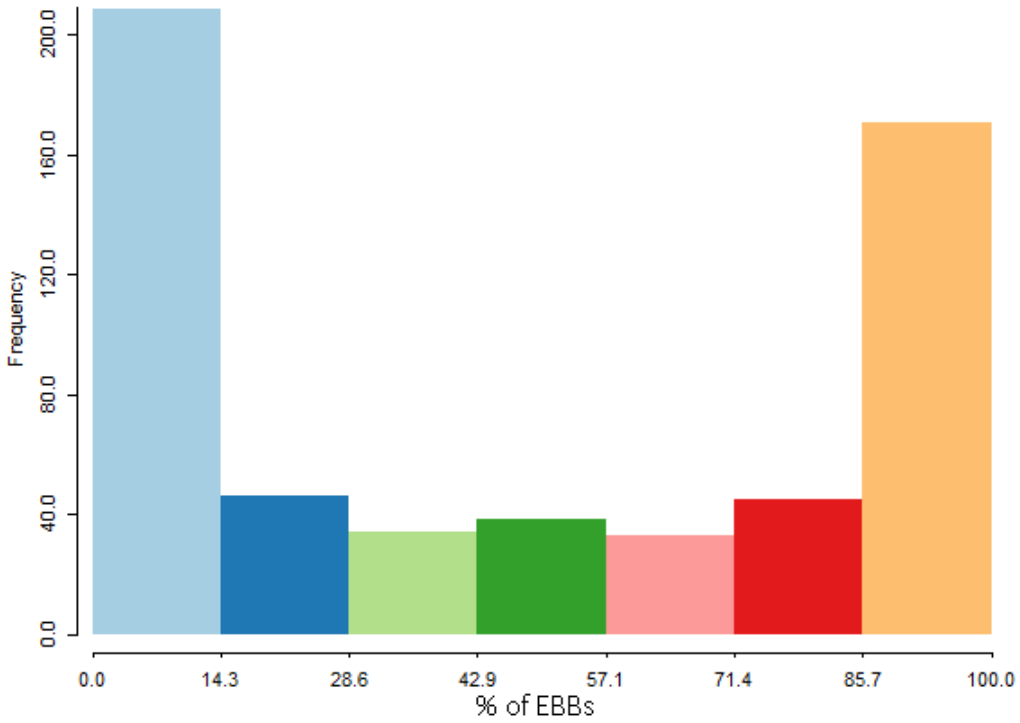
	Δ Rural Female Literacy Rate	Δ Gender Gap in Literacy
Variables	Estimate	Estimate
% of EBBs (High)	0.09***	0.03***
	(0.02)	(0.009)
% of EBBs (Low)	0.06***	0.01***
	(0.01)	(0.007)
Weighted % of EBBs (High)	-0.01	-0.003
	(0.01)	(0.007)
Weighted % of EBBs (Low)	-0.01	-0.004*
	(0.02)	(0.009)

Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

Δ Rural Female Literacy Rate: Increase in rural female literacy rate, Δ Gender Gap in Literacy: Decrease in rural gender gap in literacy rates. The above table presents the SDEM estimates for the two sub samples, low and high. The districts in the low (high) are districts with percentage of educationally backward blocks lower (higher) than the median value. The % of EBBs and the Weighted % of EBBs have the same meaning as in Tables 3.3 and 3.4, except that they are estimated for the low and high groups separately. The table shows that the SDEM estimate is larger for the high districts than the districts in the low group (.09 compared to .06). However, the association of the neighboring districts, which is measured by the Weighted % of EBBs term remains insignificant even for the high districts.

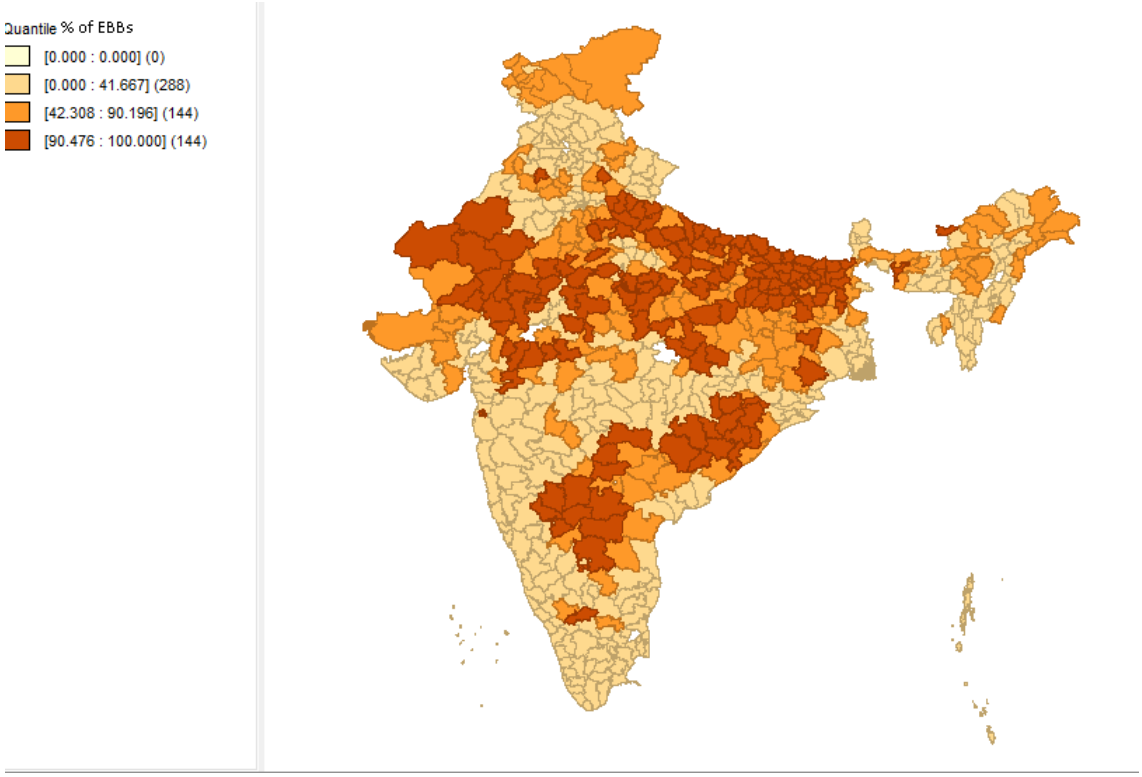
3.7 Figures

Figure 3.1: Frequency Distribution of Educationally Backward Blocks



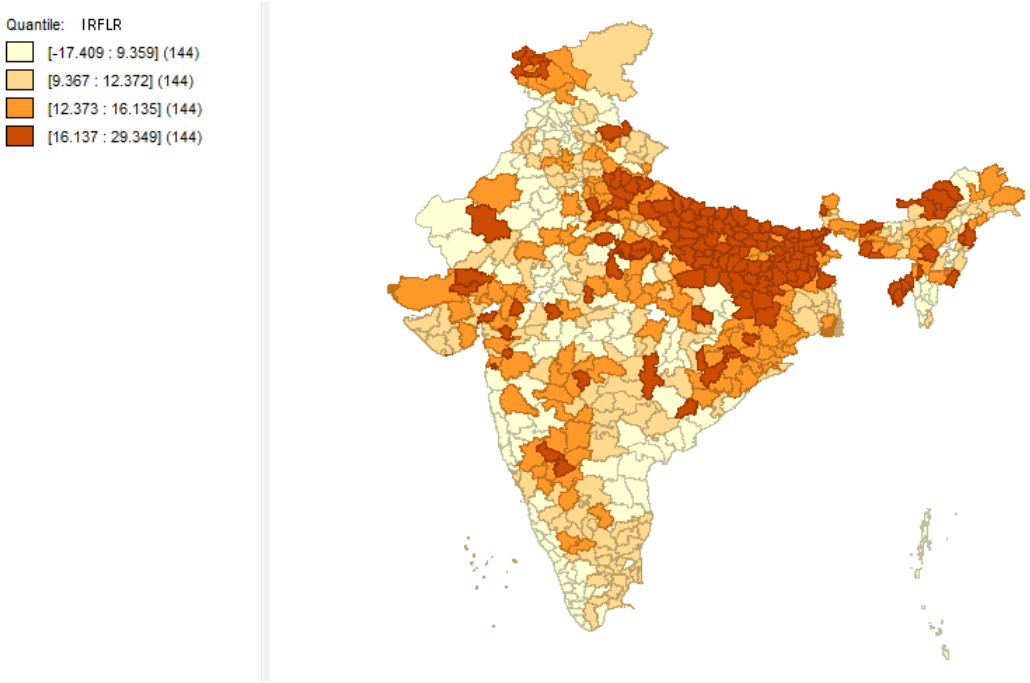
The figure presents the histogram for the percentage of educationally backward blocks (EBBs) in districts of India. There are many districts with no EBBs, whereas in some districts all the blocks are EBBs. The median percentage of EBBs is 42.9.

Figure 3.2: Distribution of Educationally Backward Blocks in Districts of India



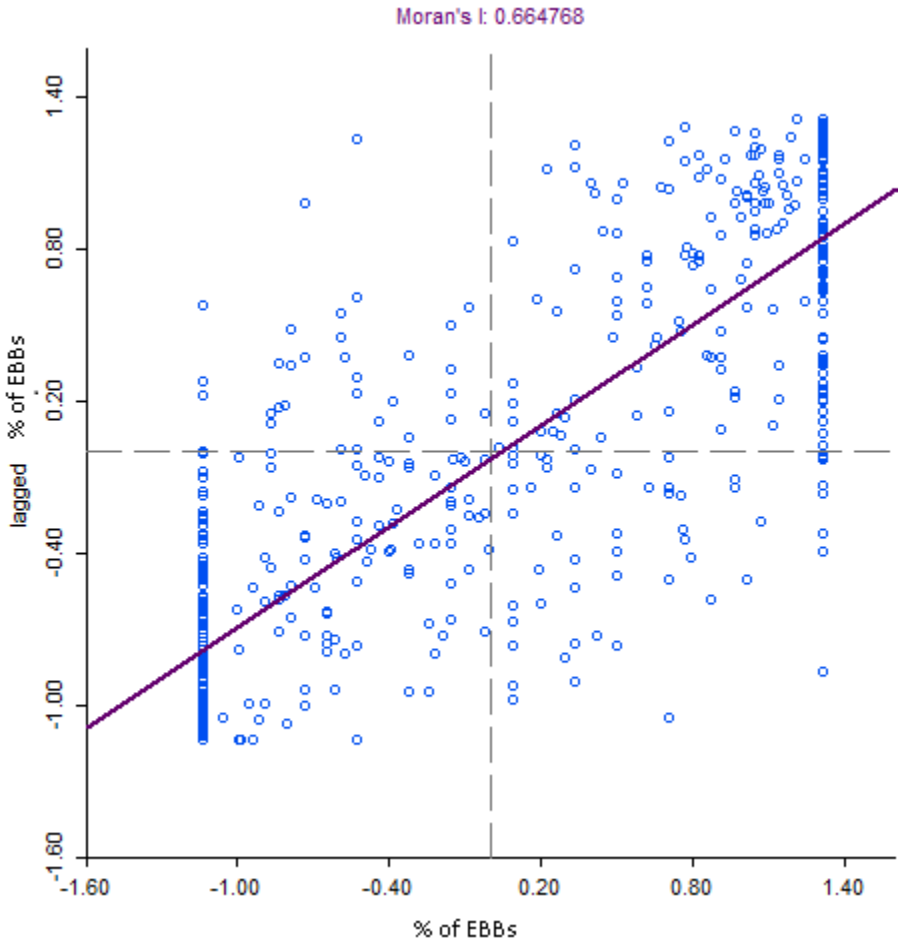
The figure shows the quantile map for the percentage of educationally backward blocks (EBBs) in districts of India. Districts with a darker shade have higher concentration of the EBBs and hence are more educationally backward than districts with a lighter shade. There is also significant concentration of the educationally backward districts in certain states of India, such as Rajasthan and Uttar Pradesh.

Figure 3.3: Quantile Map for the Outcome Variable: Increase in rural female literacy rate (IRFLR)



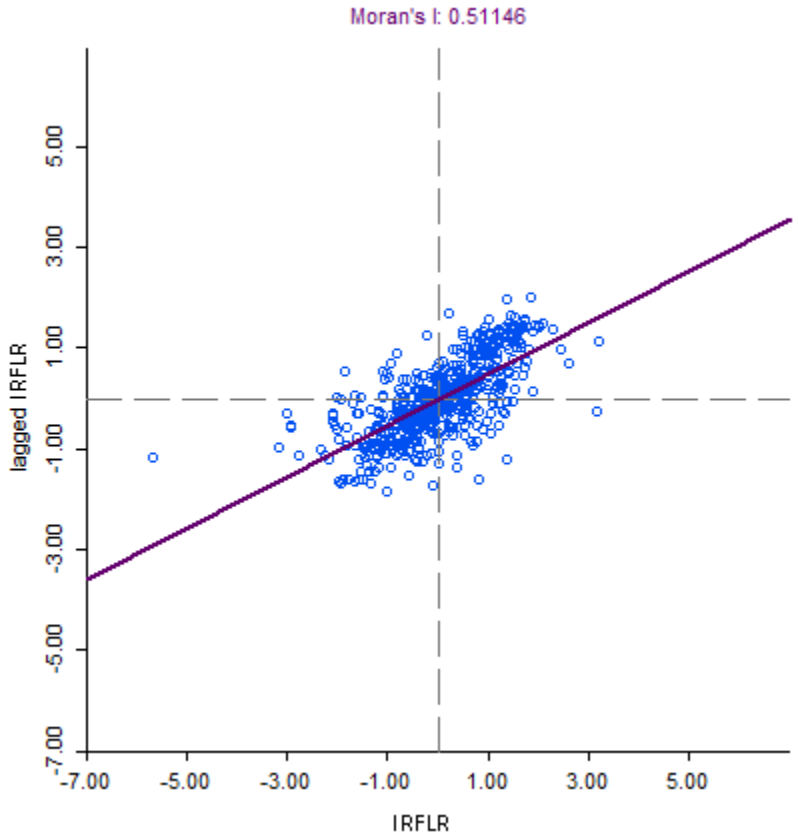
The figure shows the quantile map for the outcome variable increase in rural female literacy rates in the period of 2001-2011 in the districts of India. Districts with a darker shade have experienced a higher increase in the rural female literacy rate in the last decade. On comparing figures 3.2 and 3.3, there is some overlap of districts with high percentage of educationally backward blocks and districts that experienced a higher increase in the rural female literacy rate.

Figure 3.4: Global Moran's I plot: Percentage of Educationally Backward Blocks



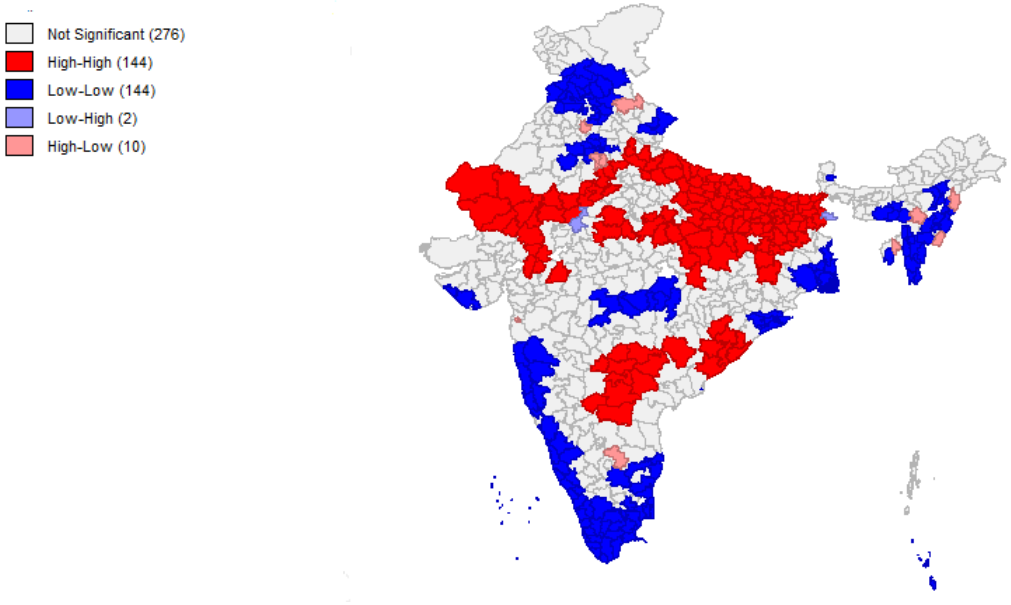
The figure presents the global Moran's I plot for the treatment variable percentage of educationally backward blocks (EBBs). The figure plots the relation between the treatment variable and it's spatial lag (that is the value of the variable for it's neighbors), The plot is constructed using standardized values on the axes, such that each unit corresponds to a standard deviation. The positively sloping fitted line through the scatter shows a positive spatial autocorrelation between the districts in the intensity of treatment. The scatter plot is centered on the mean to divide the plot into four quadrants. The top right and bottom left correspond to positive spatial autocorrelation whereas the bottom right and top left correspond to negative spatial autocorrelation.

Figure 3.5: Global Moran's I plot: Increase in rural female literacy rate (IRFLR)



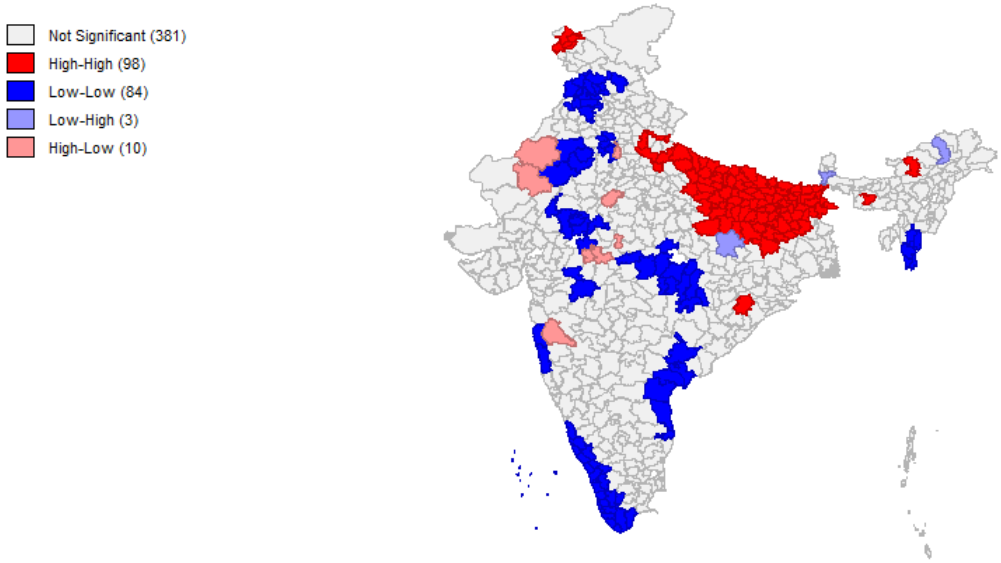
The figure presents the global Moran's I plot for the outcome variable increase in rural female literacy rate in the period of 2001-2011. The figure plots the relation between the treatment variable and its spatial lag (that is the value of the variable for it's neighbors). The plot is constructed using standardized values on the axes, such that each unit corresponds to a standard deviation and the scatter plot is centered on the mean. The positively sloping fitted line through the scatter shows a positive spatial autocorrelation between the districts in the outcome variable. The scatter plot is centered on the mean to divide the plot into four quadrants. The top right and bottom left correspond to positive spatial autocorrelation whereas the bottom right and top left correspond to negative spatial autocorrelation.

Figure 3.6: LISA cluster map: Percentage of Educationally Backward Blocks



The figure presents the local indicator of spatial autocorrelation (LISA) cluster map for the treatment variable percentage of educationally backward blocks (EBBs). High: Districts with high percentage of EBBs, Low: Districts with low percentage of EBBs. The map shows the different spatial clusters or areas where there is higher degree of spatial auto-correlation. For example a High-High region implies districts with high percentage of EBBs are surrounded by districts with high percentage of EBBs (positive spatial correlation). Similarly, the Low-Low region implies a cluster of districts with low concentration of EBBs (positive spatial correlation). A High-Low region implies negative spatial correlation as High districts are surrounded by Low districts. Not significant implies no significant local spatial autocorrelation. The significance level used to detect local spatial auto-correlation is less than 5%.

Figure 3.7: LISA cluster map: Increase in rural female literacy rate (IRFLR)



The figure presents the local indicator of spatial autocorrelation (LISA) cluster map for the outcome variable increase in rural female literacy rate (IRFLR). The map shows the different spatial clusters or hotspots, where there is higher degree of local spatial auto-correlation. The district with high (low) IRFLR are labeled as high (low). A High-High region implies districts with high IRFLR are surrounded by districts with high IRFLR (positive spatial correlation). Similarly, the Low-Low region implies a cluster of districts with low concentration of IRFLR (positive spatial correlation). A High-Low region implies negative spatial correlation as High districts are surrounded by Low districts. Not significant implies no significant local spatial autocorrelation.The significance level used to detect local spatial auto-correlation is less than 5%.

References

Afridi, F. (2010, July). Child welfare programs and child nutrition: Evidence from a mandated school meal program in India. *Journal of Development Economics 92*(2), 152–165.

Afridi, F. (2011, November). The Impact of School Meals on School Participation: Evidence from Rural India. *The Journal of Development Studies 47*(11), 1636–1656.

Aidt, T., M. A. Golden, and D. Tiwari (2011). Incumbents and Criminals in the Indian National Legislature. Technical Report 1157, Faculty of Economics, University of Cambridge.

Ajilore, O. (2011). The Impact of Ethnic Heterogeneity on Education Spending: A Spatial Econometric Analysis of United States School Districts. *Review of Urban & Regional Development Studies 23*(1), 66–76.

Akerlof, G. A. and R. E. Kranton (2000). Economics and identity. *The Quarterly Journal of Economics 115*(3), 715–753.

Anagol, S. and T. Fujiwara (2016). The Runner-Up Effect. *Journal of Political Economy 124*(4), 927–991.

Andrabi, T., J. Das, and A. I. Khwaja (2013, April). Students today, teachers tomorrow: Identifying constraints on the provision of education. *Journal of Public Economics 100*(Supplement C), 1–14.

Anselin, L. (1988). Lagrange Multiplier Test Diagnostics for Spatial Dependence and Spatial Heterogeneity. *Geographical Analysis 20*(1), 1–17.

Anselin, L. (1995). Local Indicators of Spatial Association—LISA. *Geographical Analysis 27*(2), 93–115.

Anselin, L. (2002). Under the hood Issues in the specification and interpretation of spatial regression models. *Agricultural Economics 27*(3), 247–267.

Anselin, L. (2005). Exploring Spatial Data with GeoDaTM : A Workbook. *Geographical Analysis* (1).

Anselin, L., A. K. Bera, R. Florax, and M. J. Yoon (1996, February). Simple diagnostic tests for spatial dependence. *Regional Science and Urban Economics 26*(1), 77–104.

Anselin, L., I. Syabri, and Y. Kho (2006). GeoDa : An Introduction to Spatial Data Analysis. *Geographical Analysis 38*(1), 5–22.

ASER (2011). Annual status of education report (rural).

Asher, S. and P. Novosad (2017). Politics and Local Economic Growth: Evidence from India. *American Economic Journal: Applied Economics 9*(1), 229–273.

Azam, M. and C. H. Saing (2017, November). Assessing the Impact of District Primary Education Program in India. *Review of Development Economics 21*(4), 1113–1131.

Bagde, S., D. Epple, and L. Taylor (2016). Does affirmative action work? caste, gender, college quality, and academic success in india. *American Economic Review 106*(6), 1495–1521.

Banerjee, A., D. Green, J. McManus, and R. Pande (2014). Are Poor Voters Indifferent to Whether Elected Leaders are Criminal or Corrupt? A Vignette Experiment in Rural India. *Political Communications 31*(3), 391–407.

Banerjee, A., S. Kumar, R. Pande, and F. Su (2011). Do informed voters make better choices? Experimental evidence from urban India. In *Manuscript, NBER Political Economy Meeting*. Citeseer. 00216.

Banerjee, A. and R. Somanathan (2007). The political economy of public goods: Some evidence from India. *Journal of Development Economics 82*(2), 287–314.

Banerjee, A. V., R. Banerji, E. Duflo, R. Glennerster, and S. Khemani (2010, February). Pitfalls of Participatory Programs: Evidence from a Randomized Evaluation in Education in India. *American Economic Journal: Economic Policy 2*(1), 1–30.

Banerjee, A. V., S. Cole, E. Duflo, and L. Linden (2007, August). Remedying Education: Evidence from Two Randomized Experiments in India. *The Quarterly Journal of Economics 122*(3), 1235–1264.

Banerjee, A. V. and R. Pande (2007). Parochial politics: Ethnic preferences and politician corruption. 00156.

Bardhan, P. K., D. Mookherjee, and M. Parra Torrado (2010). Impact of political reservations in West Bengal local governments on anti-poverty targeting. *Journal of Globalization and development 1*(1).

Barrera-Osorio, F., D. S. Blakeslee, M. Hoover, L. L. Linden, and D. Raju (2011). Expanding Educational Opportunities in Remote Parts of the World: Evidence from a RCT of a Public Private Partnership in Pakistan. In *Third IZA Workshop, Institute for the Study of Labor, Mexico City, Mexico*.

Benavot, A. and UNESCO (Eds.) (2015). *Education for all 2000 - 2015: achievements and challenges* (1. ed ed.). Number 12.2015 in EFA Global Monitoring Report. Paris: Unesco Publ.

Bernheim, B. D. and N. Kartik (2014). Candidates, character, and corruption. *American Economic Journal: Microeconomics 6*, 205–246.

Bertrand, M., R. Hanna, and S. Mullainathan (2010). Affirmative action in education: Evidence from engineering college admissions in india. *Journal of Public Economics 94*(1), 16–29.

Besley, T. (2005). Political selection. *The Journal of Economic Perspectives 19*(3), 43–60.

Besley, T. and A. Case (1992, March). Incumbent Behavior: Vote Seeking, Tax Setting and Yardstick Competition. Technical Report w4041, National Bureau of Economic Research, Cambridge, MA.

Besley, T., J. G. Montalvo, and M. Reynal-Querol (2011). Do Educated Leaders Matter? *The Economic Journal 121*(554), F205–F227.

Besley, T., R. Pande, and V. Rao (2005). Political selection and the quality of government: Evidence from south india.

Bhalotra, S., I. Clots-Figueras, G. Cassan, and L. Iyer (2014). Religion, politician identity and development outcomes: Evidence from India. *Journal of Economic Behavior & Organization 104*(Supplement C), 4–17.

Bhavnani, R. R. (2012). Using Asset Disclosures to Study Politicians Rents: An Application to India. In *Annual Bank Conferences on Development Economics, Washington, DC*. 00022.

Bhavnani, R. R. (2015). The effects of malapportionment on cabinet inclusion: Subnational evidence from India. *British Journal of Political Science*, 1–21. 00005.

Bird, K. (2014). Ethnic quotas and ethnic representation worldwide. *International Political Science Review 35*(1), 12–26. 00028.

Blakeslee, D. S. (2013). Politics and Public Goods in Developing Countries: Evidence from India. *Unpublished Working Paper*. 00006.

Bown, L. (1990, October). *Preparing the Future–Women, Literacy and Development. The Impact of Female Literacy on Human Development and the Participation of Literate Women in Change. ActionAid Development Report No. 4*.

Brasington, D., A. Flores-Lagunes, and L. Guci (2016, January). A spatial model of school district open enrollment choice. *Regional Science and Urban Economics 56*, 1–18.

Brueckner, J. K. (2003, April). Strategic Interaction Among Governments: An Overview of Empirical Studies. *International Regional Science Review 26*(2), 175–188.

Brueckner, J. K. and L. A. Saavedra (2001). Do Local Governments Engage in Strategic Property—Tax Competition? *National Tax Journal 54*(2), 203–229.

Burchfield, S., H. Hua, D. Baral, and V. Rocha (2002, December). *A Longitudinal Study of the Effect of Integrated Literacy and Basic Education Programs on Women's Participation in Social and Economic Development in Nepal*.

Burde, D. and L. L. Linden (2013, July). Bringing Education to Afghan Girls: A Randomized Controlled Trial of Village-Based Schools. *American Economic Journal: Applied Economics 5*(3), 27–40.

Calonico, S., M. D. Cattaneo, and R. Titiunik (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica 82*(6), 2295–2326. 00419.

Case, A. and A. Deaton (1999). School Inputs and Educational Outcomes in South Africa. *The Quarterly Journal of Economics 114*(3), 1047–1084.

Case, A. C., H. S. Rosen, and J. R. Hines (1993, October). Budget spillovers and fiscal policy interdependence: Evidence from the states. *Journal of Public Economics 52*(3), 285–307.

Chattopadhyay, R. and E. Duflo (2004a). Impact of reservation in Panchayati Raj: Evidence from a nationwide randomised experiment. *Economic and Political Weekly*, 979–986.

Chattopadhyay, R. and E. Duflo (2004b). Women as Policy Makers: Evidence from a Randomized Policy Experiment in India. *Econometrica 72*(5), 1409–1443. 01254.

Chemin, M. (2012). Welfare Effects of Criminal Politicians: A Discontinuity-Based Approach. *The Journal of Law and Economics 55*(3), 667–690.

Chin, A. (2005). Can redistributing teachers across schools raise educational attainment? Evidence from Operation Blackboard in India. *Journal of Development Economics 2*(78), 384–405.

Chin, A. and N. Prakash (2011). The redistributive effects of political reservation for minorities: Evidence from India. *Journal of Development Economics 96*(2), 265–277. 00067.

Clots-Figueras, I. (2011). Women in politics: Evidence from the Indian States. *Journal of Public Economics 95*(7–8), 664–690. 00137.

Clots-Figueras, I. (2012). Are Female Leaders Good for Education? Evidence from India. *American Economic Journal: Applied Economics 4*(1), 212–244.

Costa, H., L. G. Veiga, and M. Portela (2015, September). Interactions in Local Governments' Spending Decisions: Evidence from Portugal. *Regional Studies 49*(9), 1441–1456.

Dall'erba, S. (2005, March). Distribution of regional income and regional funds in Europe 1989–1999: An exploratory spatial data analysis. *The Annals of Regional Science 39*(1), 121–148.

Das, J., S. Dercon, J. Habyarimana, P. Krishnan, K. Muralidharan, and V. Sundararaman (2013). School Inputs, Household Substitution, and Test Scores. *American Economic Journal: Applied Economics 5*(2), 29–57.

Duflo, E. (2001, September). Schooling and Labor Market Consequences of School Construction in Indonesia: Evidence from an Unusual Policy Experiment. *American Economic Review 91*(4), 795–813.

Duflo, E., R. Hanna, and S. P. Ryan (2012, June). Incentives Work: Getting Teachers to Come to School. *American Economic Review 102*(4), 1241–1278.

Dunning, T. and J. Nilekani (2013). Ethnic Quotas and Political Mobilization: Caste, Parties, and Distribution in Indian Village Councils. *American Political Science Review 107*(1), 35–56.

Duraisamy, P. and B. Jérôme (2017). Who wins in the Indian parliament election: Criminals, wealthy and incumbents? *Journal of Social and Economic Development*, 1–18.

Dutta, Bhaskar and Gupta, P. (2015). How Indian Voters Respond to Candidates with Criminal Charges. *Economic and Political Weekly 2014 49, 50*.

Echávarri, R. A. and R. Ezcurra (2010, February). Education and gender bias in the sex ratio at birth: Evidence from India. *Demography 47*(1), 249–268.

Epple, D., R. Romano, and H. Sieg (2008). Diversity and Affirmative Action in Higher Education. *Journal of Public Economic Theory 10*(4), 475–501.

Ferraz, C. and F. Finan (2008). Exposing corrupt politicians: The effects of brazil's publicly released audits on electoral outcomes. *Quarterly Journal of Economics 123*(2), 703–745.

Filmer, D. and N. Schady (2008). Getting girls into school: Evidence from a scholarship program in Cambodia. *Economic development and cultural change 56*(3), 581–617.

Fisman, R., F. Schulz, and V. Vig (2017). Financial disclosure and political selection: Evidence from India.

Fredriksson, P. G. and D. L. Millimet (2002a, November). Is there a 'California effect' in US environmental policymaking? *Regional Science and Urban Economics 32*(6), 737–764.

Fredriksson, P. G. and D. L. Millimet (2002b, January). Strategic Interaction and the Determination of Environmental Policy across U.S. States. *Journal of Urban Economics 51*(1), 101–122.

Gehring, K., T. F. Kauffeldt, and K. C. Vadlamannati (2015). Crime, Incentives and Political Effort: A Model and Empirical Application for India. Technical Report 170, Courant Research Centre PEG.

Ghosh, S. (2010). Strategic interaction among public school districts: Evidence on spatial interdependence in school inputs. *Economics of Education Review*, 440–450.

Girard, V. (2016). Mandated political representation and crimes against the low castes. *World Institute for Development Economic Research (UNU-WIDER) Working Paper* (074).

Glewwe, P., M. Kremer, and S. Moulin (2009, January). Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics 1*(1), 112–135.

Glewwe, P., M. Kremer, S. Moulin, and E. Zitzewitz (2004, June). Retrospective vs. prospective analyses of school inputs: the case of flip charts in Kenya. *Journal of Development Economics 74*(1), 251–268.

Glick, P. (2008, September). What Policies will Reduce Gender Schooling Gaps in Developing Countries: Evidence and Interpretation. *World Development 36*(9), 1623–1646.

Gonzalez Canche, M. S. (2014, December). Localized competition in the non-resident student market. *Economics of Education Review 43*, 21–35.

Greenbaum, R. T. (2002, March). A spatial study of teachers' salaries in Pennsylvania school districts. *Journal of Labor Research 23*(1), 69–86.

Holzer, H. and D. Neumark (1999). Are Affirmative Action Hires Less Qualified? Evidence from Employer--Employee Data on New Hires. *Journal of Labor Economics 17*(3), 534–569.

Htun, M. (2004). Is Gender like Ethnicity? The Political Representation of Identity Groups. *Perspectives on Politics 2*(3), 439–458.

Imbens, G. and K. Kalyanaraman (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *The Review of Economic Studies 79*(3), 933–959.

Iyer, L. and A. Mani (2011). Traveling Agents: Political Change and Bureaucratic Turnover in India. *The Review of Economics and Statistics 94*(3), 723–739.

Iyer, L., A. Mani, P. Mishra, and P. Topalova (2012). The Power of Political Voice: Women's Political Representation and Crime in India. *American Economic Journal: Applied Economics 4*(4), 165–193.

Iyer, L. and M. Reddy (2013). Redrawing the Lines: Did Political Incumbents Influence Electoral Redistricting in the World's Largest Democracy. Technical report, Harvard Business School Working Paper. 00012.

Jacoby, H. G. (2002). Is There an Intrahousehold "Flypaper Effect"? Evidence From a School Feeding Programme. *Economic Journal 112*(476), 196–221.

Jalan, J. and E. Glinksya (2013, August). Improving primary school education in India : an impact assessment of DPEP - phase one. Technical Report 81375, The World Bank.

Jensenius, F. R. (2015). Development from representation? A study of quotas for the scheduled castes in India. *American Economic Journal: Applied Economics 7*(3), 196–220. 00019.

Jogani, C. (2018). Does More Schooling Infrastructure affect Literacy.

Kazianga, H., D. Levy, L. L. Linden, and M. Sloan (2013, July). The Effects of "Girl-Friendly" Schools: Evidence from the BRIGHT School Construction Program in Burkina Faso. *American Economic Journal: Applied Economics 5*(3), 41–62.

Khanna, G. (2015). Large-scale education reform in general equilibrium: Regression discontinuity evidence from india. *Working Paper, University of Michigan*.

Kremer, M., E. Miguel, and R. Thornton (2009, August). Incentives to Learn. *Review of Economics and Statistics 91*(3), 437–456.

Krishnan, N. (2007). Political reservations and rural public good provision in india. *Boston University*. 00028.

Lahoti, R. and S. Sahoo (2017). Are Educated Leaders Good for Education? Evidence from India. SSRN Scholarly Paper ID 2851895, Social Science Research Network, Rochester, NY.

Lee, D. S. and T. Lemieux (2010). Regression discontinuity designs in economics. *Journal of Economic Literature 48*(2), 281–355.

Linden, L. (2004). Are incumbents really advantaged? The preference for non-incumbents in Indian national elections. *Unpublished paper*.

Lloyd, C., C. Mete, and Z. Sathar (2005, April). The Effect of Gender Differences in Primary School Access, Type, and Quality on the Decision to Enroll in Rural Pakistan. *Economic Development and Cultural Change 53*(3), 685–710.

Lundberg, J. (2006, August). Spatial interaction model of spillovers from locally provided public services. *Regional Studies 40*(6), 631–644.

McCrary, J. (2008). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics 142*(2), 698–714.

McMillen, D. P., L. D. Singell, and G. R. Waddell (2007). Spatial Competition and the Price of College. *Economic Inquiry 45*(4), 817–833.

Meller, M. and S. Litschig (2015, November). Adapting the Supply of Education to the Needs of Girls: Evidence from a Policy Experiment in Rural India. *Journal of Human Resources*.

MHRD (2009). Sarva sikshya abhiyan framework for implementation. *Department of School Education and Literacy*.

Millimet, D. L. and T. Collier (2008, July). Efficiency in public schools: Does competition matter? *Journal of Econometrics 145*(1), 134–157.

Min, B. and Y. Uppal (2011). Estimating the Effects of Quotas Across India using Satellite Imagery. 00000.

Misra, K., P. W. Grimes, and K. E. Rogers (2012, December). Does competition improve public school efficiency? A spatial analysis. *Economics of Education Review 31*(6), 1177–1190.

Mukhopadhyay, B. (2014). Elections in India: Strategic Nominations. *Review of Market Integration 6*(1), 8–46.

Muralidharan, K. (2017). Chapter 3 - Field Experiments in Education in Developing Countries. In A. V. Banerjee and E. Duflo (Eds.), *Handbook of Economic Field Experiments*, Volume 2 of *Handbook of Economic Field Experiments*, pp. 323–385. North-Holland.

Muralidharan, K. and N. Prakash (2017, July). Cycling to School: Increasing Secondary School Enrollment for Girls in India. *American Economic Journal: Applied Economics 9*(3), 321–350.

Murray, R. (2015). What Makes a Good Politician? Reassessing the Criteria Used for Political Recruitment. *Politics & Gender 11*(04), 770–776.

Nath, A. (2015). Bureaucrats and Politicians: How Does Electoral Competition Affect Bureaucratic Performance? *Institute for Economic Development (IED) Working Paper 269*. 00016.

Nichols, A. (2016, September). RD: Stata module for regression discontinuity estimation.

Paisa (2012). Do schools get their money. *Paisa Report, An Accountability Initiative*.

Pande, R. (2003). Can Mandated Political Representation Increase Policy Influence for Disadvantaged Minorities? Theory and Evidence from India. *The American Economic Review 93*(4), 1132–1151. 00476.

PEO (2010). Evaluation report on sarva sikshya abhiyan. *Program Evaluation Organisation, Planning Commission of India*.

Prakash, N., M. Rockmore, and Y. Uppal (2014). Do Criminally Accused Politicians Affect Economic Outcomes? Evidence from India. Technical Report 192, Households in Conflict Network.

Robinson-Pant, A. (2006). The social benefits of literacy. *Documento de referencia para el Informe de Seguimiento de la EPT en el Mundo*.

Singh, A., A. Park, and S. Dercon (2014). School Meals as a Safety Net: An Evaluation of the Midday Meal Scheme in India. *Economic Development and Cultural Change 62*(2), 275 – 306.

UNDP (2012). Empowering Women for Stronger Political Parties.

UNESCO (2015). Promising practices in the asia-pacific region. *UNESCO Case Study on India*.

Uppal, Y. (2009). The disadvantaged incumbents: estimating incumbency effects in Indian state legislatures. *Public Choice 138*(1), 9–27.

Vaishnav, M. (2012). The merits of money and muscle: Essays on criminality, elections and democracy in india. *Columbia University*.

Vega, S. H. and J. P. Elhorst (2015). The Slx Model. *Journal of Regional Science 55*(3), 339–363.

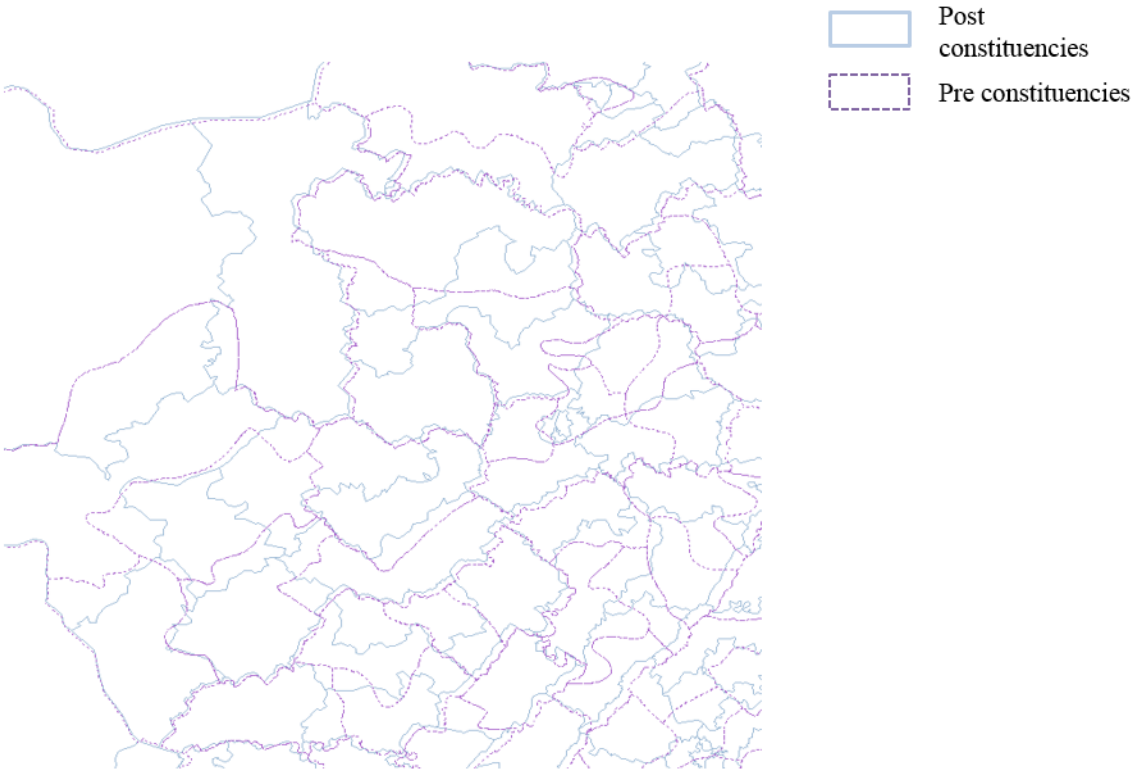
Wilkinson, S. I. (2006). The politics of infrastructural spending in India. *Department of Political Science, University of Chicago, mimeo 31*.

Appendix A

APPENDICES TO CHAPTER 1

A.1 Additional Figures and Tables

Figure A.1: Change in boundaries of constituencies due to latest Redistricting

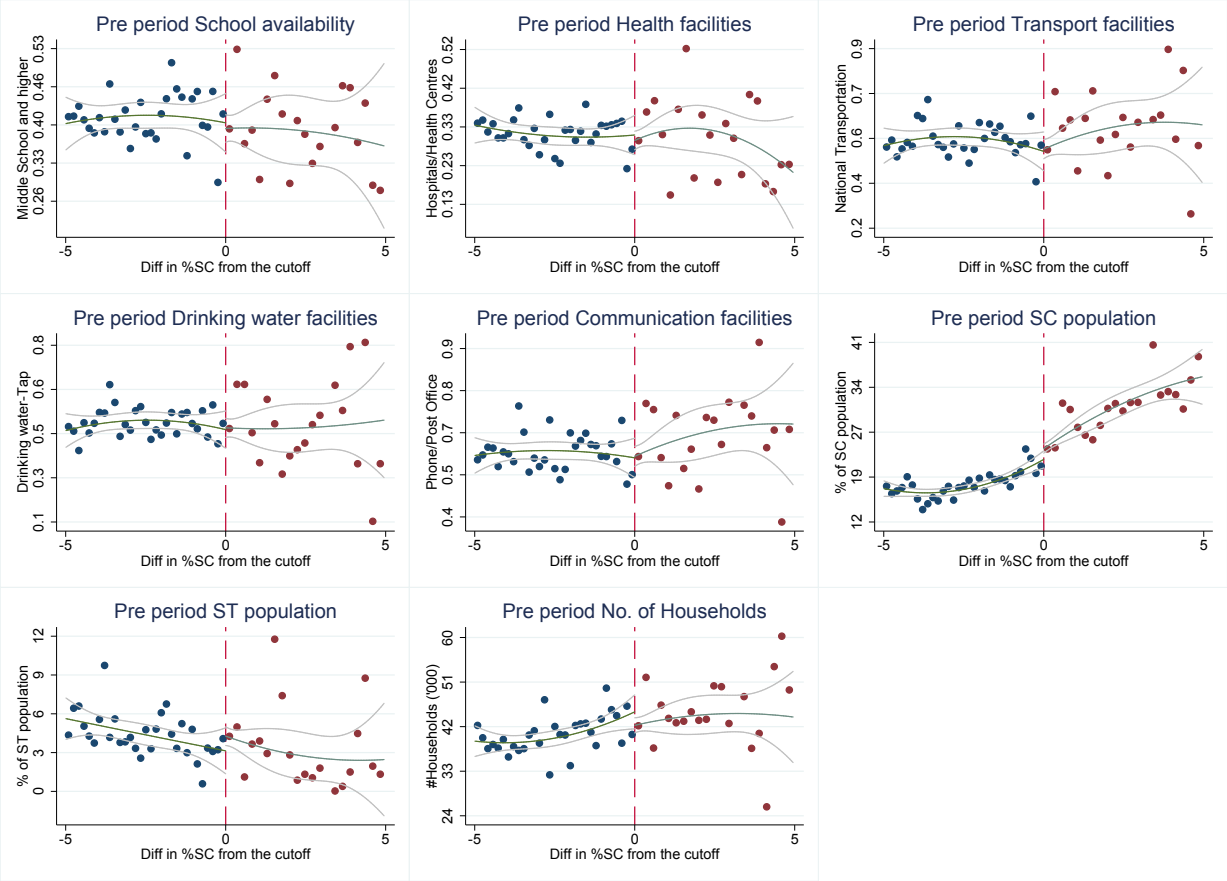


A sample of the boundaries of assembly constituencies before and after the latest redistricting which occurred after a gap of three decades. There seems to be a significant change in the boundaries of the constituencies.

Figure A.2: Skipping a Spatially Adjacent Constituency for SC Reservation

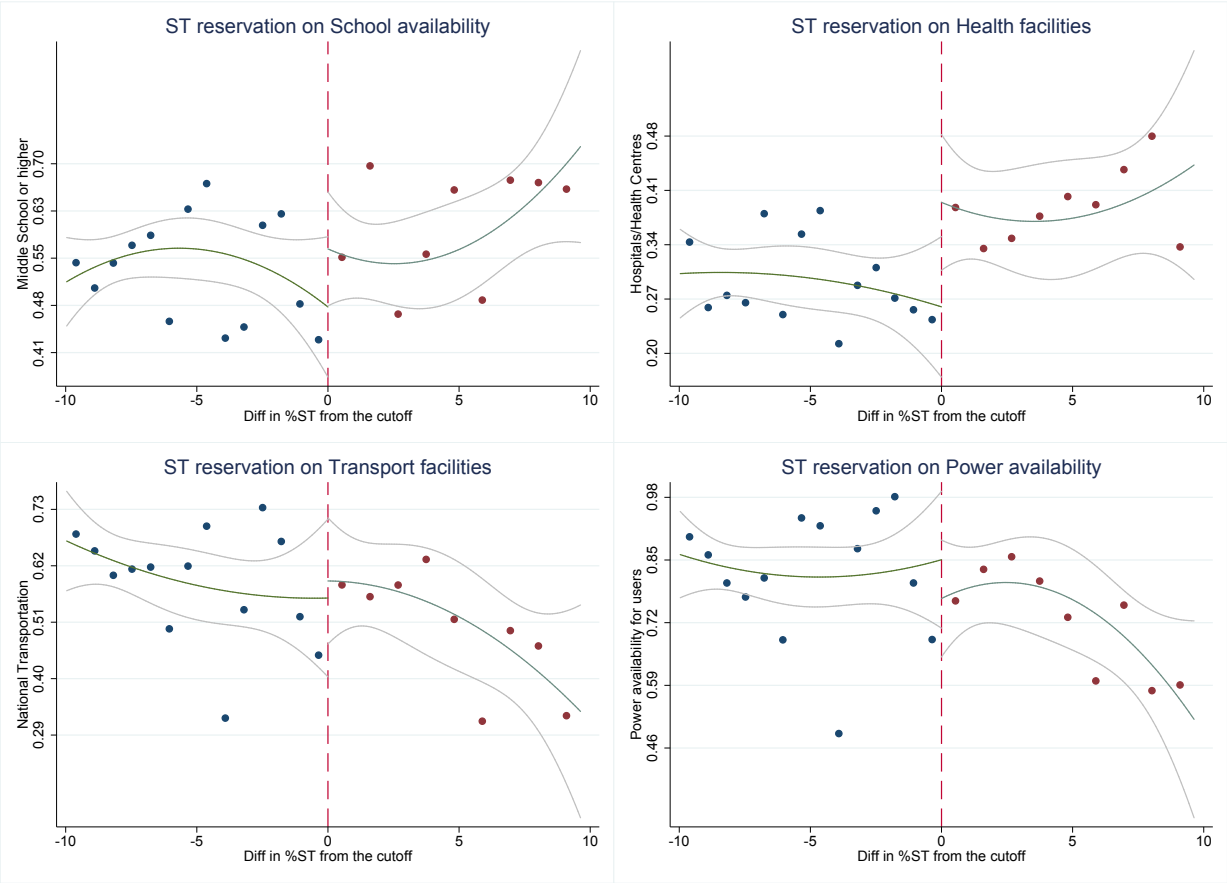
Constituency Addanki although eligible for reservation was skipped and Yerragondapalem was reserved instead. This was because Addanki was adjacent to Santhanuthalapadu which had the highest percentage of Scheduled Castes population in Prakasam district in the state of Andhra Pradesh. Source of the figure: http://www.cmsir.com/tdp/mla-profiles/. The figure has been modified slightly for presentation.

Figure A.3: Covariates Balance Test for RD validity



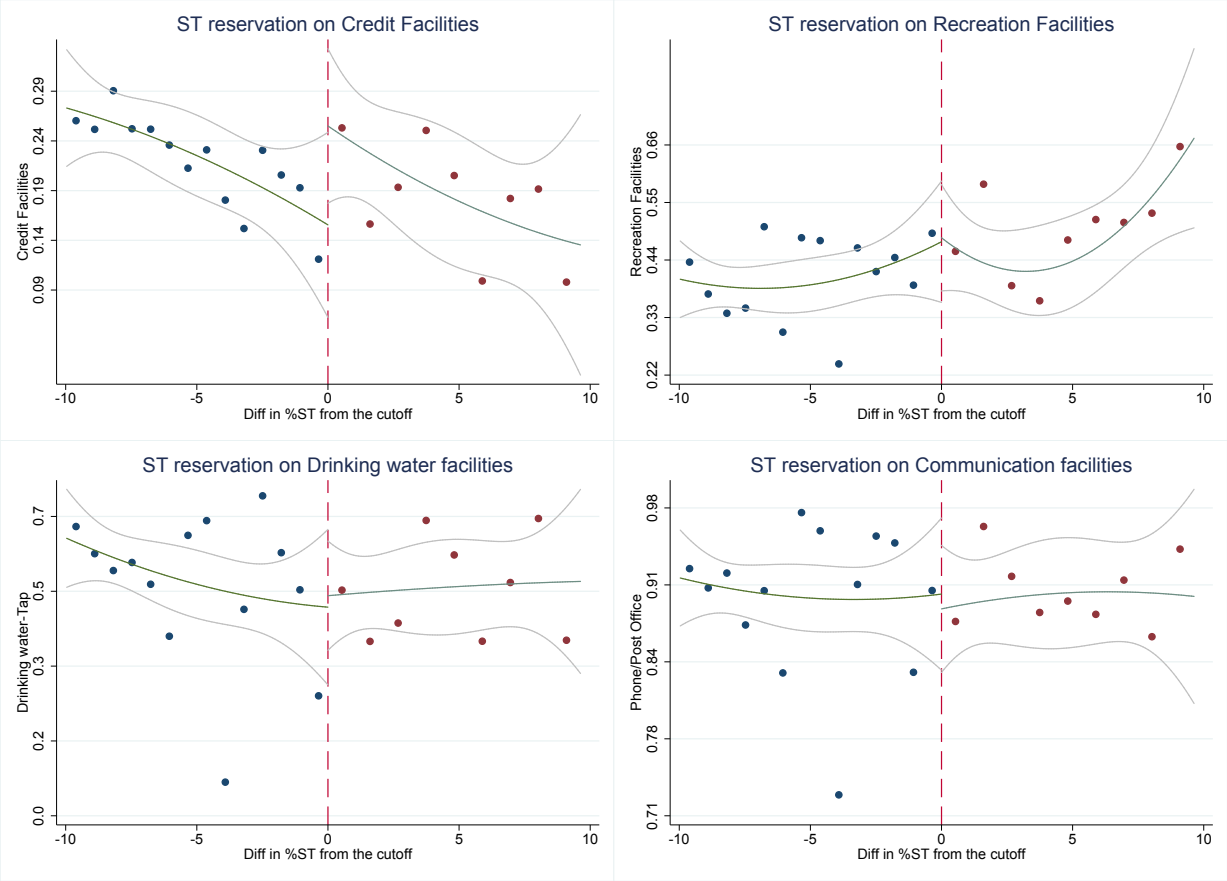
There was no discontinuity at the cutoff for any of the covariates.

Figure A.4: Effect of ST reservation on Level of Village Facilities



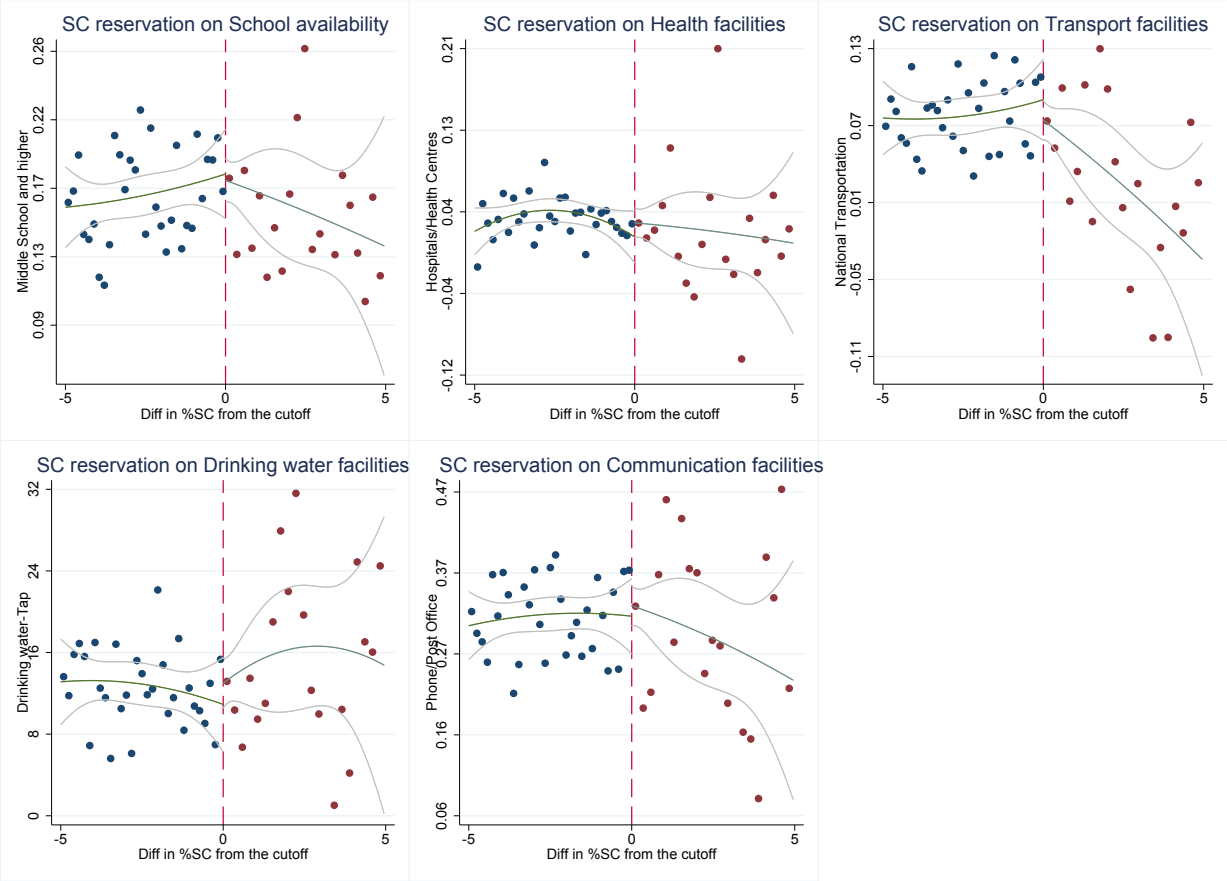
The confidence intervals are large due to small number of constituencies reserved for Scheduled Tribes. There does not seem to be a difference across reservation status of constituencies. Some positive effect on health facilities seems plausible.

Figure A.5: Effect of ST reservation on level of village facilities



The figure does not suggest a difference between the levels of recreation, drinking water and communication facilities across unreserved and reserved constituencies. However, not many villages seem to have credit facilities and the availability decreases for constituencies with higher percentage of Scheduled Tribes population.

Figure A.6: Effect of SC reservation on growth of village Facilities



The outcome variables represent the growth in the level of village facilities from 2001 to 2011. The variables have been restricted to those that could be found in both the Censuses. There has been a positive increase in all the facilities, but the SC constituencies seem to be falling behind in case of transportation facilities.

Table A.1: Controlling for number of years from elections to Census

VARIABLES	(1) Middle School or higher	(2) Health Centers	(3) Transport	(4) Electricity	(5) Credit facilities	(6) Recreation facilities	(7) Tap	(8) Phone/Post Office
Reserved for SC	-0.029	-0.019	-0.033	-0.001	0.023	-0.015	-0.046	-0.013
	(0.035)	(0.039)	(0.056)	(0.016)	(0.047)	(0.056)	(0.059)	(0.016)
No. of Observations	530	549	532	550	501	534	537	550

Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The table presents estimates for constituencies which had a gap of more than 2 years from the state elections to the Census. The results remain similar and insignificant. The number of observations indicate the number of constituencies within the optimal CCT bandwidth.

A.1.1 Attributes of Candidates contesting in National Elections

In this section I explore whether there is difference in attributes of candidates for the national elections due to reservation of parliamentary constituencies. A parliamentary constituency (PC) is the relevant electoral unit for national elections. Delimitation defines boundaries of PCs and there are 543 PCs. The national elections are first past the post system; candidate with the highest vote in the PC is a Member of Parliament. The Delimitation Commission also reserves PCs for the Scheduled Castes and Scheduled Tribes using a similar algorithm as for the assembly constituencies. I use a similar strategy of establishing a discontinuous relation between the proportion of reserved population and reservation status.

Reservation of PCs for Scheduled Castes include similar exceptions as reservation of assembly constituencies for Scheduled Castes, such as maintaining heterogeneity in the geographic distribution of SC constituencies. As shown in the figure below, the probability of reservation for SC increases by approximately 50 percentage points on crossing the threshold. Since, the number of PCs are much smaller compared to assembly constituencies, there is a higher proportion of PCs affected due to the exception for SC reservation.

Table A.2: Reservation of PCs

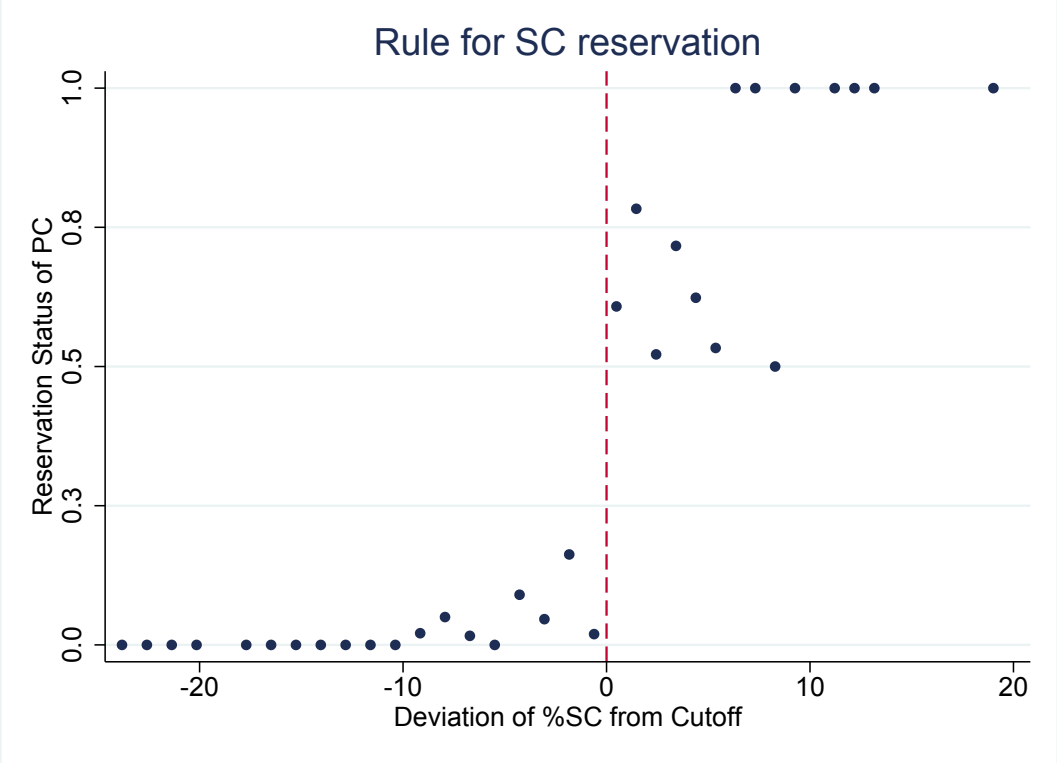
VARIABLES	(1) Reserved for SC
RD estimate	0.522***
	(0.105)
Observations	2,521
BW	CCT
Control	% of SC Population

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

The number of observations is equal to the number of candidates within a bandwidth of 5.3 percentage points around the cutoff of zero. The estimates remain similar for different bandwidths. Percentage of Scheduled Castes in the PC has been used as control.

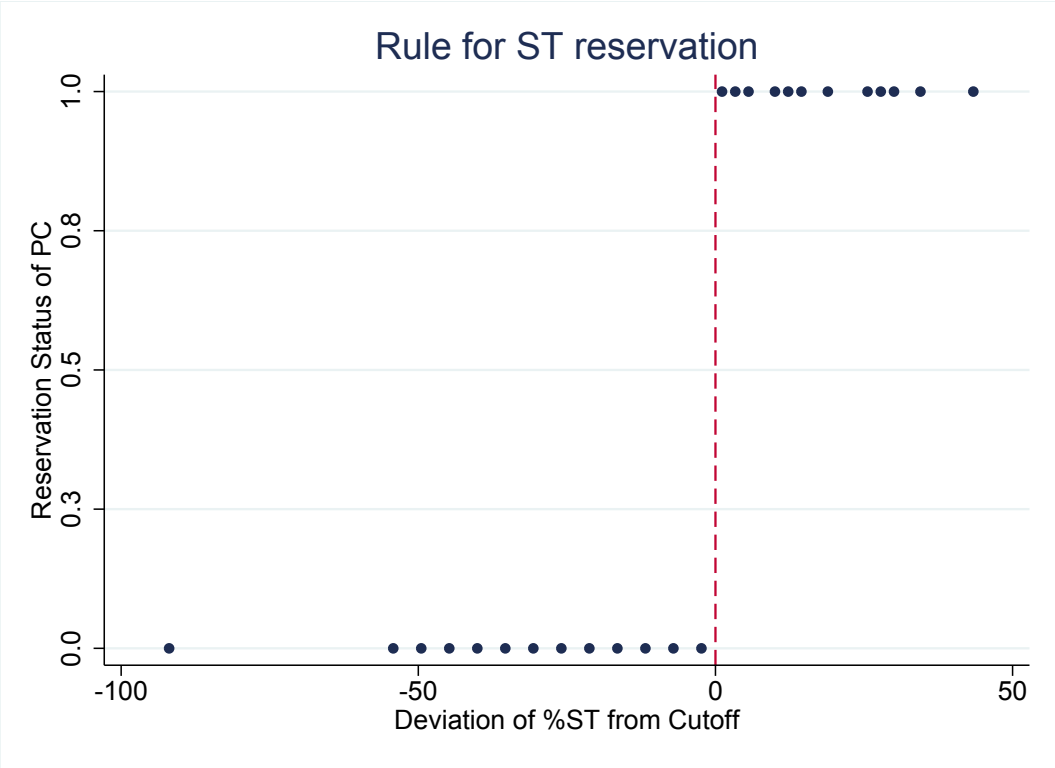
Figure A.7: Reservation of Parliamentary Constituencies for SCs



The percentage population of Scheduled Castes corresponding to the last PC reserved acts as the cutoff for this figure. The running variable is normalised to have the cutoff as zero. All other points are differences of the percentage population from the cutoff.

There are no exceptions to reservation of PCs for Scheduled Tribes. As shown in the figure below, a PC is reserved with a probability of one on crossing the population cutoff for Scheduled Tribes.

Figure A.8: Reservation of Parliamentary Constituencies for ST



The cutoff is normalized to zero and thus all other points are differences of the percentage population from the cutoff.

For PCs too, these imply a fuzzy and sharp regression discontinuity design. But, the sample size is insufficient to find non parametric second stage RD estimates. Instead, I use two stage least-squares method with a specification similar to that of fuzzy RD design and obtain parametric estimates. I restrict the sample to a bandwidth of ten percentage points. I use affidavits of candidates for national elections after the latest redistricting, that is in 2009 and 2014. The table below presents the results for the effect of reservation of a PC on attributes of candidates. Candidates from SC reserved constituencies have lower criminal records and total assets. The result are same on controlling for incumbents. The estimates for total assets are negative but insignificant due to high standard errors.

A.1.2 Overlapping old and new constituencies

I find the overlap area between the old and new constituencies by using geocoded maps of constituencies. I overlay maps of the old constituencies on the new constituencies and find the overlap

Table A.3: Effect on SC Reservation on Attributes of Candidates

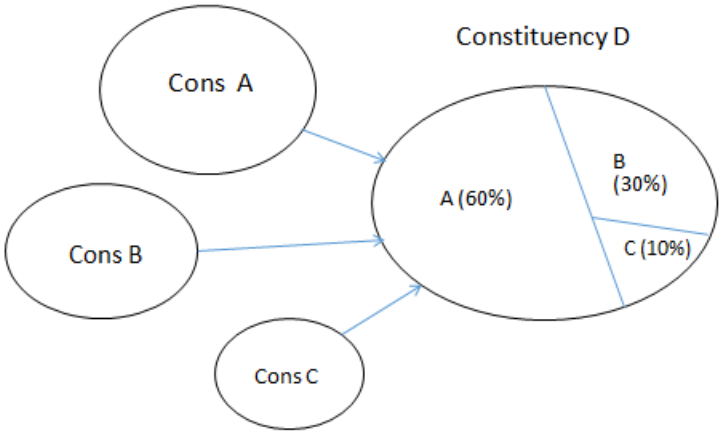
VARIABLES	(1) Has a criminal record	(2) Total Assets	(3) No.of Criminal Cases
Reserved for SC	-0.119***	-7.299	-0.363**
	(0.0389)	(4.720)	(0.152)
Constant	0.127***	33.45***	0.270
	(0.0400)	(5.198)	(0.170)
Observations	5,611	5,548	5,610
R-squared	0.036	0.105	0.023
Control	Y	Y	Y

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

The number of observations imply the number of candidates within a bandwidth of ten percentage points, cutoff normalized to zero. The nature of the estimates remain similar for different bandwidths. The regressions control for percentage of Scheduled Castes and include state fixed effects.

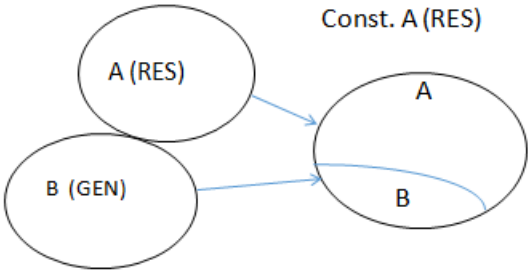
percentage using the software ArcGIS. For every new constituency, I find the old constituency with the largest area in the new constituency. The diagram below illustrates this:



The new constituency D comprises of 60 percent of the old constituency A, 30 percent of B, and 10 percent of C. The old constituency approximately similar to new constituency D is A. In case of a 100 percent overlap this would imply that the old and new constituencies are the same. Else, I find the constituency that is most similar to the new constituency.

A.1.3 Effect of a Change in Reservation Status on Individual Villages

The strategy can be better understood using the following picture:



Let us assume the old constituencies A and B form the new constituency A. A was a reserved constituency both before and after the redistricting, whereas B was not. The villages in B that now belong to A, faced a change in constituency as well as a change in reservation status (of the constituencies they belong to). To identify villages that experienced the change, I need to know the villages that belonged to the old and new constituencies. I acquire the mapping between the villages and the old constituencies from the data used in Jensenius (2015). Raphael Susewind shared the mapping of villages with the new constituencies and I update the data wherever required. I combine both datasets and construct dummy variables for: whether a village experienced a change in constituency, and whether the village experienced a change in reservation status.

Thus, based on whether a village experienced a change in boundary or reservation status, all the villages can be classified as shown in the following table.

	Reservation Status Change		
Boundary Change	Gained Reservation	Lost Reservation	No Changers
Yes			
No			

The reservation status of a village could change because it switched constituencies, or because the reservation status of the constituency it belonged to changed.[1] I estimate the change in the level of public goods in a village because of the change in reservation status of the constituency using

[1] I use the initial sample of 2,801 constituencies and the 401,431 villages that comprised them. There were 14,410 villages for which the sample did not have the old reservation status and 2,166 observations for which it was unclear if the villages changed constituencies. This leads to a final sample of 384,855 villages used in this analysis.

the following specification:

$$\Delta Y_i = \alpha + \beta_1 SW_i + \beta_2 BC_i + \beta_3 SW_i * BC_i + \Delta Pop_r + \lambda_c + \varepsilon_i \tag{A.1}$$

$$SW_i = \begin{cases} 1 \text{ if General to Reserved} \\ 0 \text{ if General to General} \end{cases} \tag{A.2}$$

where ΔY_i is the change in the level of village facilities from 2001 to 2011, SW_i=1 implies village i is reserved after the redistricting and earlier was not, BC_i indicates if the village changed constituencies.[2] The coefficient of interest is β_3 for the interaction term $SW_i * BC_i$ which captures the effect of the change in reservation status from general to reserved for villages that changed constituencies. ΔPop_r is the growth in the reserved population comprising of Scheduled Castes and Scheduled Tribes in the village from 2001 to 2011. The specification includes the old constituency fixed effect denoted by λ_c and the errors are clustered at the constituency level.

Result from the above regression is presented in the table below and the estimates obtained are insignificant and negative. The coefficients imply that on gaining a reservation status, percentage of villages that have a middle school or higher decreases by 1.6 percentage points. Incorporating the confidence interval, this implies any negative effect of greater than 2.8 percentage points can be ruled out. This is not very different from the estimates obtained from the regression discontinuity estimation for constituencies.[3]

[2]This implies that the control group in this case are villages which were unreserved before and after the redistricting. I also estimate a specification using all villages that did not change reservation status including those that were reserved both before and after the redistricting as control group. The results do not change significantly.

[3]I also validate results by considering SC and ST reservation separately, the results do not change significantly.

Table A.4: Effect on individual villages

VARIABLES	(1) Middle School or Higher	(2) Tap	(3) Phone/Post Office	(4) Hospitals/Health Centers	(5) Transport
Interact	-0.0156	-0.0151	-0.0178	0.00375	-0.00685
	(0.0123)	(0.0225)	(0.0221)	(0.0145)	(0.0137)
GEN to Reserved	0.00676	0.0198	0.000884	0.00271	0.00307
	(0.0115)	(0.0178)	(0.0186)	(0.0104)	(0.0106)
Boundary Change	0.00410	-0.0156*	0.00290	0.000725	-0.00816*
	(0.00393)	(0.00804)	(0.00824)	(0.00478)	(0.00470)
Constant	0.129***	0.178***	0.371***	0.0130	0.0498
	(0.0242)	(0.0455)	(0.0601)	(0.0261)	(0.0418)
Observations	348,862	347,207	347,200	342,502	346,297
R-squared	0.017	0.048	0.033	0.022	0.014
Control	Y	Y	Y	Y	Y
Fixed Effects	Y	Y	Y	Y	Y

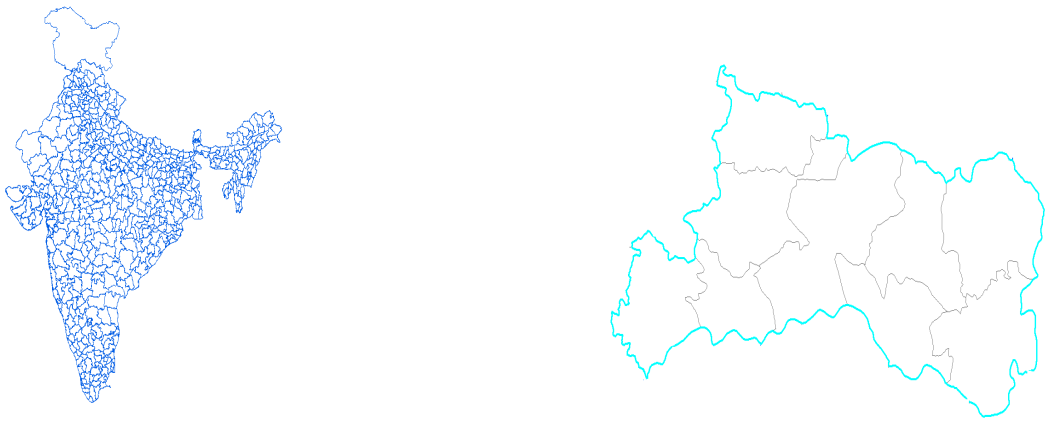
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

All the specifications include constituency level fixed effects and the growth in reserved population as control. The standard errors have been clustered at the constituency level. The estimates of interest are presented in the first row of the table.

A.1.4 Effect of Reservation of Assembly constituencies on Facilities in a District

An analysis of the effect of reservation of assembly constituencies on a district may help to account for any spillovers between constituencies and internalize any friction due to change in boundary of the constituencies. The assembly constituencies are never split between districts and contained in one district. The below figure presents the map of districts in India and assembly constituencies in a sample district:

Figure A.9: Districts and Assembly Constituencies in India

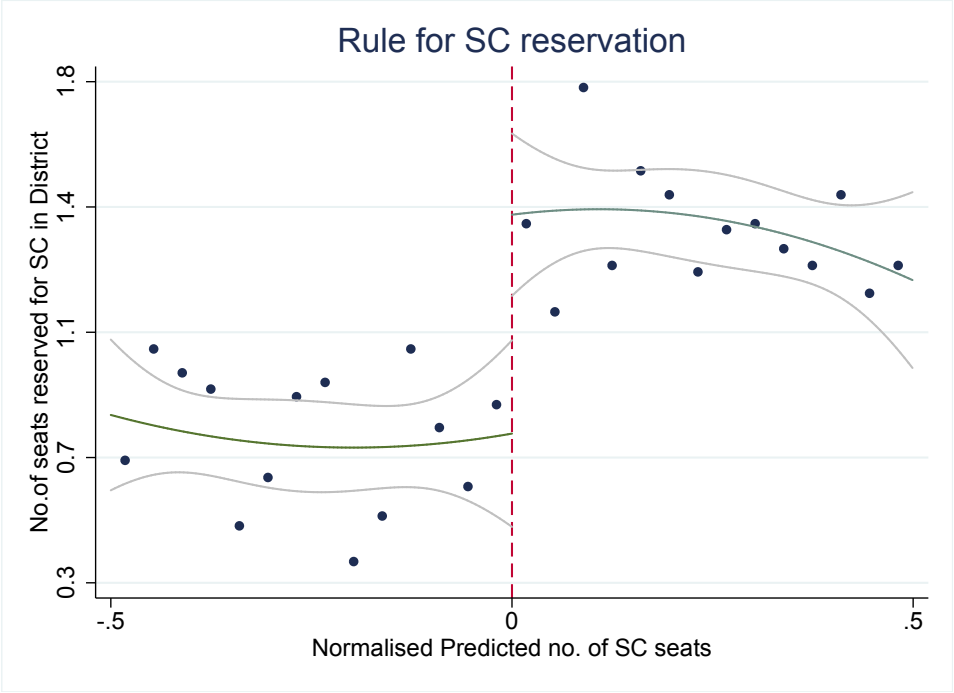


(a) Districts (b) ACs in a District

The question now relevant is slightly different, what is the effect of an increase in the number

of reserved constituencies in a district. I use two methods to analyze this question. Figure 2 shows that the number of constituencies reserved for Scheduled Castes in a district will change depending on if the predicted number of constituencies is greater than the 0.5 thresholds (that is 0.5, 1.5, 2.5 and so on). I stack all the thresholds and normalize them to be zero such that observations with predicted number of seats between zero to 0.5 are on the left of the cutoff.

Figure A.10: Allocation of SC constituenciss to districts



This leads to a RD setup as shown in the above figure, where the number of SC seats or constituencies is a discontinuous function of the predicted number of SC seats. The first stage estimation for the relation is presented in the table below:

Table A.5: First stage estimates

VARIABLES	(1) #SC seats	(2) #SC seats
RD Estimate	0.426**	0.480***
	(0.204)	(0.178)
Observations	452	452
Band Width	.17	.35

*** p<0.01, ** p<0.05, * p<0.1

The table shows that the probability of getting one more SC seat increases by 0.43 for a district on crossing the cutoff. The optimal bandwidth is estimated using the CCT procedure and is 0.17

But, due to a small sample size, it is challenging to obtain precise estimates for a given bandwidth. The RD plots for the second stage does not show any significant effect.

For the district level, I also explore another variation. Due to the latest redistricting, there was a change in the number of assembly constituencies in a district and also a change in the number of reserved assembly constituencies. So, some districts gained or lost the number of reserved assembly constituencies. I construct a panel setup and estimate it using the following specifications:

$$Y_{it} = \alpha + \alpha_1 SCG_i + \alpha_3 Post_i + \alpha_4 Post_i * SCG_i + \varepsilon_i \tag{A.3}$$

where SCG_i: increase in number of SC seats in district i and $Post_i$=1 for 2011

Another alternative specification would be as follows:

$$Y_{it} = \alpha + \alpha_1 SCG_i + \alpha_2 SCL_i + \alpha_3 Post_i + \alpha_4 Post_i * SCG_i + \alpha_5 Post_i * SCL_i + \varepsilon_i \tag{A.4}$$

where SCG_i: increase in number of SC seats in district i, SCL_i: decrease in number of SC seats in district i, and $Post_i$=1 for 2011

The results remain insignificant and do not differ significantly in either of the specifications.

A.2 Data Appendix

A.2.1 Algorithm based on the Procedure of Reservation of Constituencies

The total number of assembly constituencies (ACs) in India has remained constant at 4,120. The number of reserved constituencies depends on the relative share of the reserved group and may change with the population growth. The total number of constituencies reserved for Scheduled Castes (SC) or Scheduled Tribes (ST) can be derived using the below formula:

$$\text{Total No. of ACs reserved for SC (ST)} = \frac{\text{Total Population of SC (ST) in India* 4120}}{\text{Total Population of India}}$$

The number of constituencies allocated to a state depends on the population share of the state:[4]

$$\text{No. of ACs in a State} = \frac{\text{Population of state * 4120}}{\text{Total Population of India}}$$

and the number of constituencies reserved in each state is based on the population share of the reserved group, that is

$$\text{No. of ACs reserved for SC/ST in a State} = \frac{\text{State Population of SC/ST*No. of ACs in the State}}{\text{Total Population of the State}}$$

For example, consider the state of Madhya Pradesh. The population numbers according to the 2001 census is as below:

2001 Population= 60 million
2001 SC Population= 9 million
2001 ST Population= 12 million

The total number of ACs allocated to the state is 230 out of 4120.
No. of ACs reserved for SC in a State $= \frac{9*230}{60}$=34.89=35.

Likewise, the number of ACs reserved for ST in a State is 41.

The Delimitation Commission divides the state into constituencies with similar population levels.

[4]But, the 2001 population was not used to allocate constituencies across states. Thus, the number of constituencies allocated to states remained the same in the latest delimitation.

To maintain geographic heterogeneity for SC constituencies within a state, there is allocation of SC constituencies (seats) across districts. The number of predicted SC constituencies for a district can be derived using the following formula:

$$\text{Pred No. of SC seats in a District (X)} = \frac{\text{District Population of SC*No. of SC seats in the State}}{\text{Total Population of SC in the State}}$$

From the formula above, the predicted number of constituencies, or X, can be a fraction. Since the number of constituencies cannot be a fraction, the Commission uses the following criteria to obtain the number of SC constituencies for a district:

Range of X	SC seats
0-0.5	0
0.5-1.5	1
1.5-2.5	2
2.5-3.5	3

To illustrate further, the state of Madhya Pradesh has 48 districts and 35 SC constituencies. Let us consider the district of Sheopur in Madhya Pradesh: Sheopur has approximately one percent of the SC population of the state, hence the predicted number of SC constituencies for Sheopur = $.01 * 35$ =.35. Since $.35 < .5$, the number of SC constituencies allocated to Sheopur is zero.

But, the district Morena has 3.6 percent of the SC population of the state, hence the predicted number of SC constituencies for Morena = $.036 * 35 = 1.28$. Since $1.28 < 1.5$, the number of SC constituencies allocated to Morena is one. For the district Sagar, the predicted number is 1.59 which is greater than 1.5. Therefore, the district Sagar is allocated two constituencies as reserved for SC. Figure 2 represents the underlying step function used for allocation of SC constituencies to districts.

After the allocation of SC constituencies to a district, the Commission reserves constituencies with the highst SC population in the district. Hence, out of the six constituencies in the district Morena, the constituency Ambah with the highest SC population is reserved. Whereas, for the district Sagar, the constituencies Naryoli and Bina, with the highest and second higest SC population are reserved. For ST reservation, there is no extra step of allocation across districts. Thus, the Commission reserves the 41 ACs with the highest ST population in Madhya Pradesh for ST.

A.2.2 Level of Public Goods for Constituencies

The census data on village facilities does not provide the constituency a village belongs to. Thus, I required a mapping between the villages and constituencies to obtain the provision of public goods for the constituencies. Since, I compare outcomes before and after the latest redistricting too, I needed the mapping of villages to both the old and new constituencies. For this I use data from two different sources.

Mapping Villages to Old constituencies

I use the data submitted on the American Economic Journal website Jensenius (2015). I would like to acknowledge the source of this data. See Jensenius (2015) for details of the data. I use this data to determine if a village belonged to a different constituency before and after the redistricting or experienced a change in reservation status. I also use the data for some robustness checks.

Mapping Villages to New constituencies

To study the effect of quota on provision of public goods, I find the level of village facilities in 2011 for the new constituencies. The mapping between the villages in 2011 and the new constituencies was shared generously by Raphael Susewind. The data is protected under the Open Data Commons Open Database license. This dataset was created using proprietary data of the village location coordinates and shapefiles of the new constituencies. Some of the mapping was incorrect because the shapefiles used were not accurate enough for a large scale analysis. The villages in this data had the village codes of Census 2001.

To verify the consistency and accurateness of the data, I use information on the administrative areas of the constituencies. The Delimitation commission has reports of the redistricting process that specifies the administrative areas of the constituencies. The administrative areas of the constituencies specified are district, blocks, sub-blocks (Revenue Inspection circles and Patwari Circles). Unfortunately, the villages that comprise a constituency is not provided. I use this information on the extent of the constituencies as a starting point for verifying and updating the data provided by Raphael Susewind. The information on the extent of the constituencies was also compiled by S Anand and shared on Datameet. But, this data did not have the census codes for most of the administrative divisions. Hence, I had to match by names of the administrative divisions.

Methodology followed for checking the accuracy of the data

The lowest level of administrative unit in the delimitation report for the extent of the constituencies was not consistent across all states. For example, some constituencies reported the extent in terms of sub-blocks, but not others. To maintain consistency I considered extent of the constituencies in terms of blocks, which was available for all states. I aggregate the data to have constituency-block pairs. This would provide the correct composition of the constituencies. Similarly, I aggregate the data on village-constituencies to have constituency-block pairs using the mapping of villages and new constituencies. I match these two data sets to single out the pairs in our data that do not exist in the original delimitation reports. This implies that if there were villages mapped to an incorrect constituency, the particular constituency-block pair would not exist in the papers of the Delimitation Commission. These helped weed out some villages that were mapped incorrectly.

Another way to spot the erroneous cases was that the wrong constituency-block pairs would have very few villages. I used this procedure to filter the correct cases from the incorrect or doubtful ones. I could not match all constituency-block pairs across these two data sets due to matching using strings. I updated the correct constituency names for the incorrect pairs manually from the delimitation papers. I was able to do so for approximately 99 percent of the villages. I also used another incomplete mapping between villages and constituencies shared by the state of Madhya Pradesh to update the constituencies for some of the villages in it.

Public goods data at the village level for 2001 and 2011

Having established the mapping between villages of 2001 and the new constituencies, the next step was to map this to the 2011 data. The village codes are different for 2001 and 2011. To link the data I requested the directory of the census codes from the Census division of India. Few villages split or merged with other villages between 2001 and 2011. To tackle this issue I aggregate the 2011 village facilities at the 2001 census level. This results in public good variables for 2001 and 2011 corresponding to villages that existed in 2001.

A.3 Other Notes

1. At the state level, for majority of the states, the legislatures are composed of the Governor and only one house called the legislative assembly or Vidhan Sabha. The states of Bihar, Jammu and Kashmir, Karnataka, Maharashtra and Uttar Pradesh have an additional house called the legislative council or Vidhan Parishad, similar to the Rajya Sabha at the national level. Andhra Pradesh and the newly formed state Telangana have the legislative council since 2007 and 2014. Among the union territories only Delhi and Pondicherry have legislative assemblies, thus other union territories are excluded from my analysis.

2. State legislators are also responsible for appointing members of the upper house of Parliament of India called Rajya Sabha along with the President of India. The president can appoint 12 members to be exact on the basis of exceptional contribution or expertise in variious fields, such as science and art. (Source: rajyasabha.nic.in)

3. It is responsible for authorizing state expenditures, borrowing and taxes. The power to originate money bills rests solely with the Legislative assemblies. Sales tax and VAT are the sources of income for state governments. (Source: knowindia.gov.in)

4.To give an idea about how assembly constituencies overlap with the administrative divisions, there are 35 states and union territories, 593 districts, 5,143 blocks in India according to Census 2001 and 4,120 assembly constituencies.

5.There is availability of several datasets for the administrative districts in India, but extrapolating those variables for the assembly constituencies would be an inaccurate approximation.

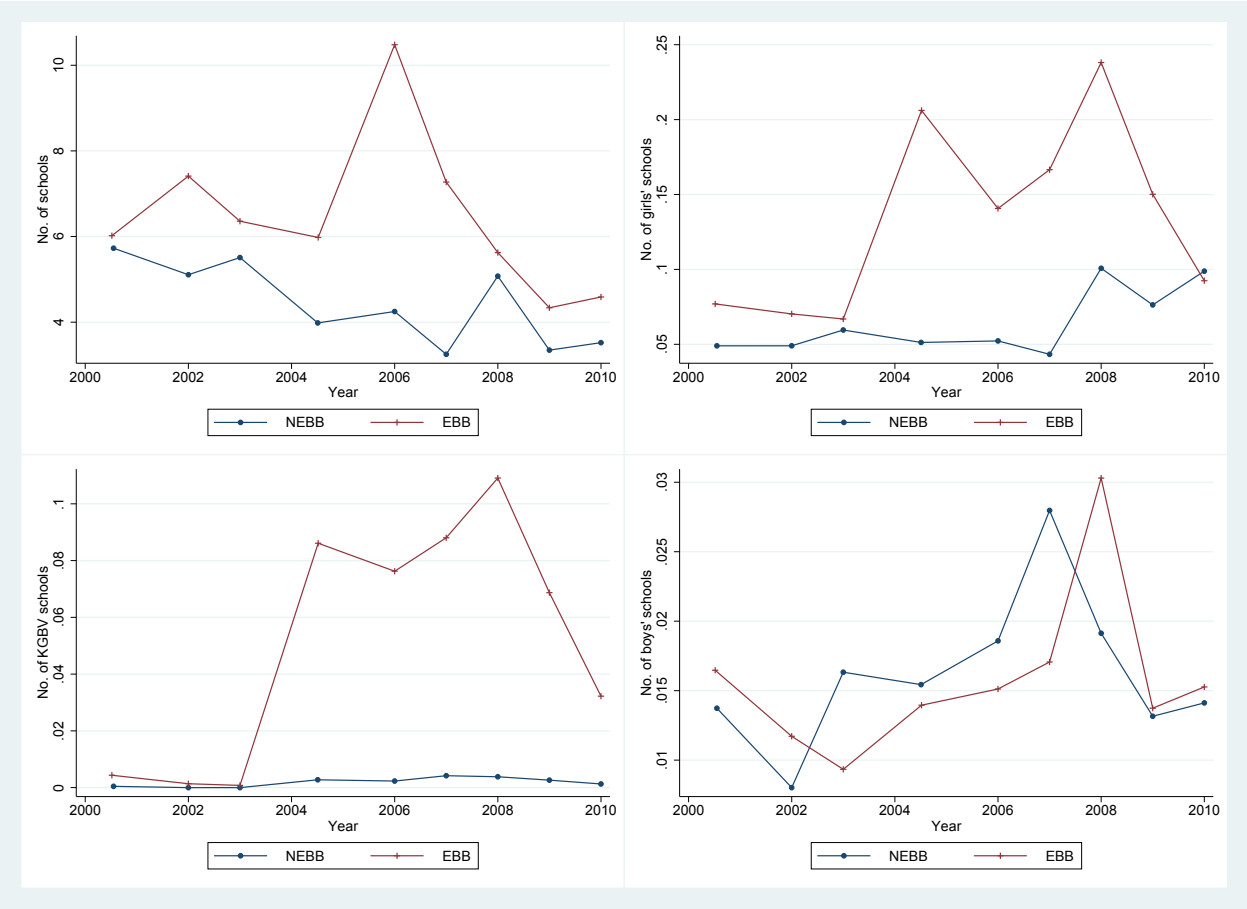
6. The Representation of People's Act, 1951 bars convicted citizens from contesting elections, but there is no law against candidates with criminal charges. The Act of 1951 has some criticisms as there might be bias in prosecuting the elites or the powerful with charges against them to prevent or delay conviction making them eligible to stand for elections. An amendment to this Act has been proposed, but not implemented yet (Dutta, 2015).

Appendix B

APPENDICES TO CHAPTER 2

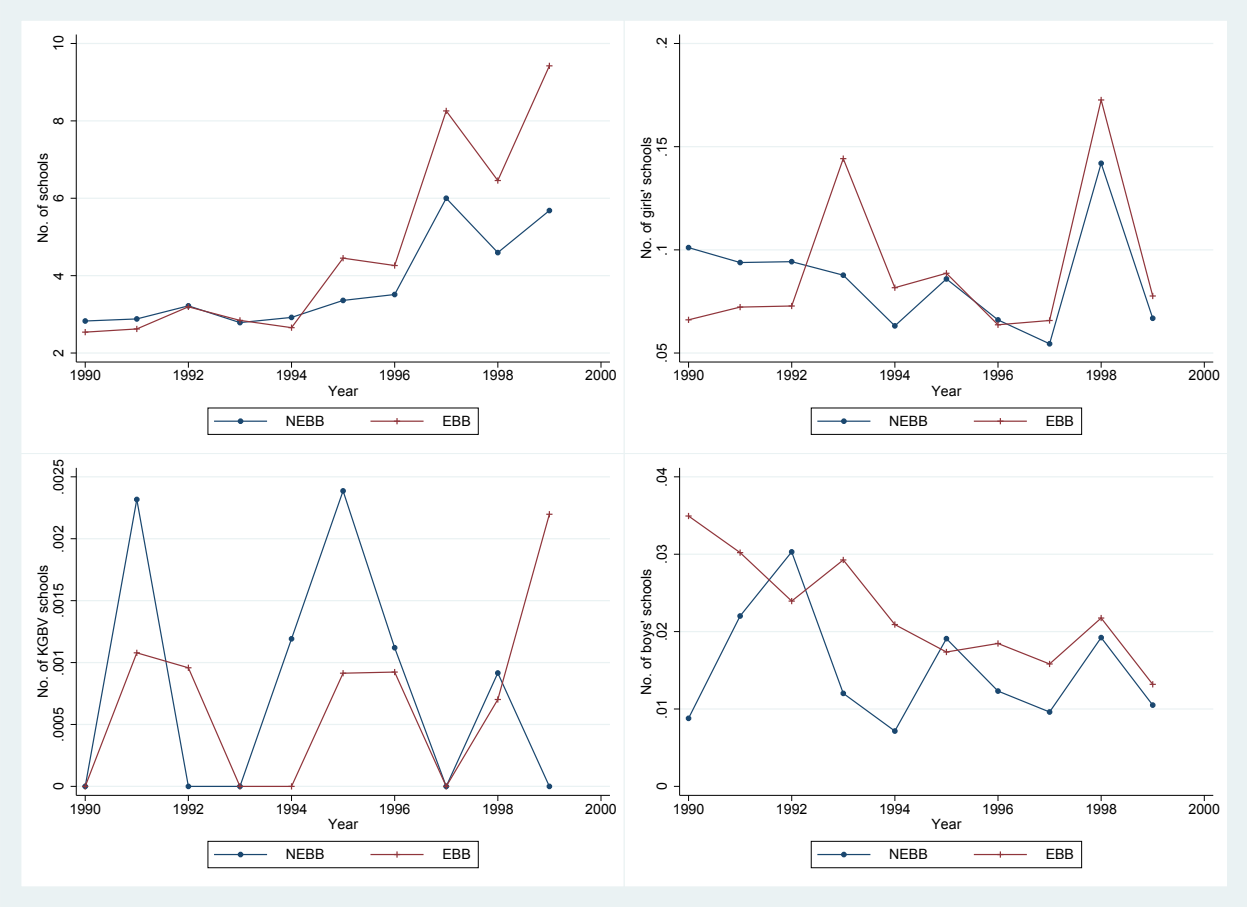
B.1 Additional Figures

Figure B.1: Average number of Schools built in the last decade



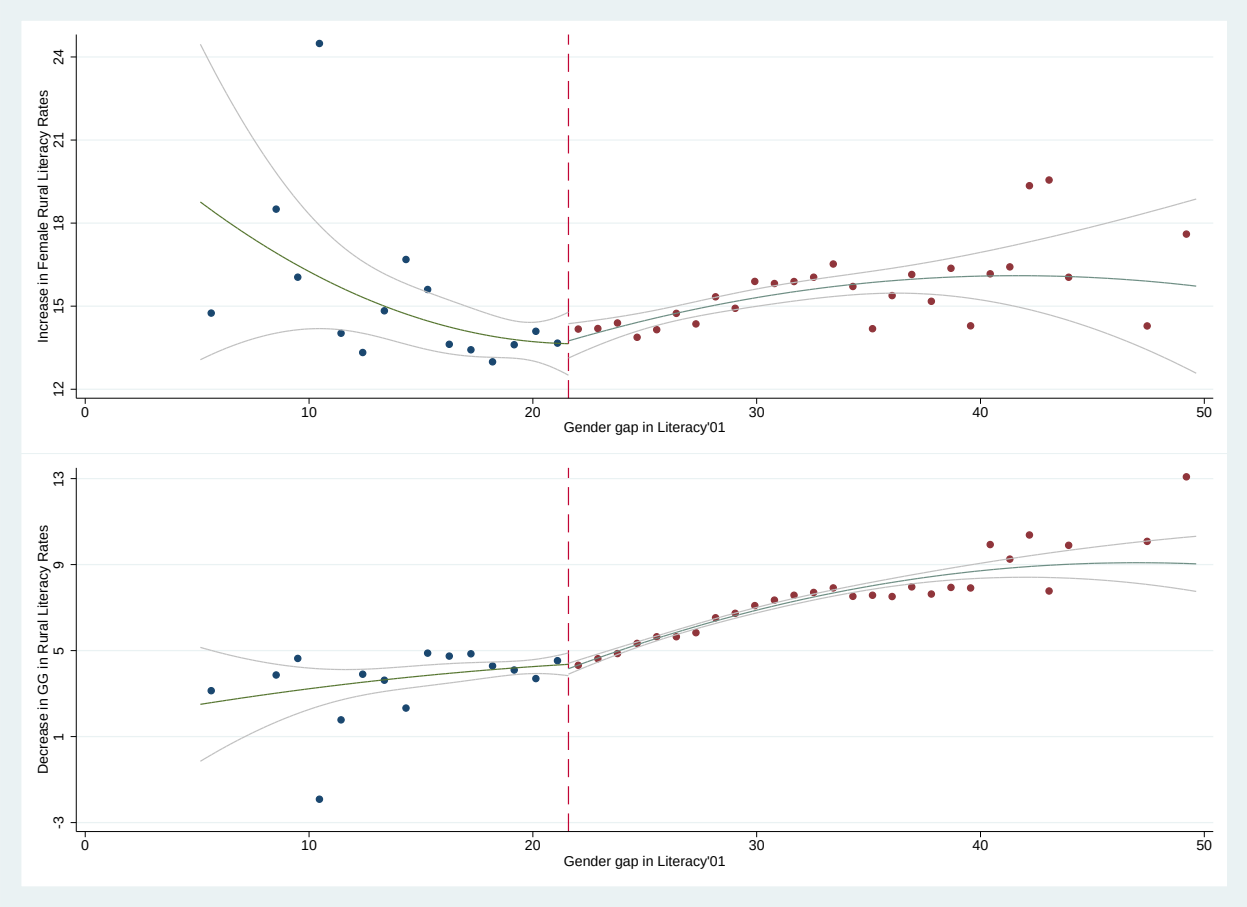
The figure plots the average number of schools built in an EBB/NEBB in the last decade. The growth of different kinds of schools is also depicted above. We see that in the past decade on average more schools were built in an EBB compared to a NEBB. KGBV schools have been built only in the EBBs and the number of girls' schools have been on the rise too.

Figure B.2: Average number of Schools built in the years 1990-2000



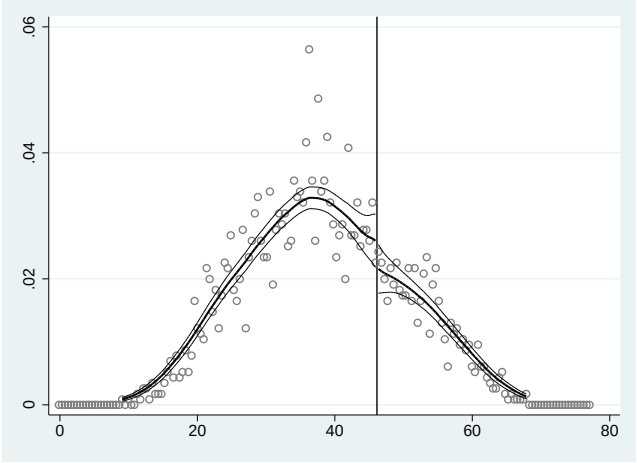
The figure plots the average number of schools built in an EBB/NEBB in the decade of 1990-2000. The growth of different kinds of schools is also depicted above. We see the number of schools built in an EBB vs a NEBB is not very different in this decade. According to the data there were fifteen KGBV schools in total that were built in this decade.

Figure B.3: RD on switching the criteria



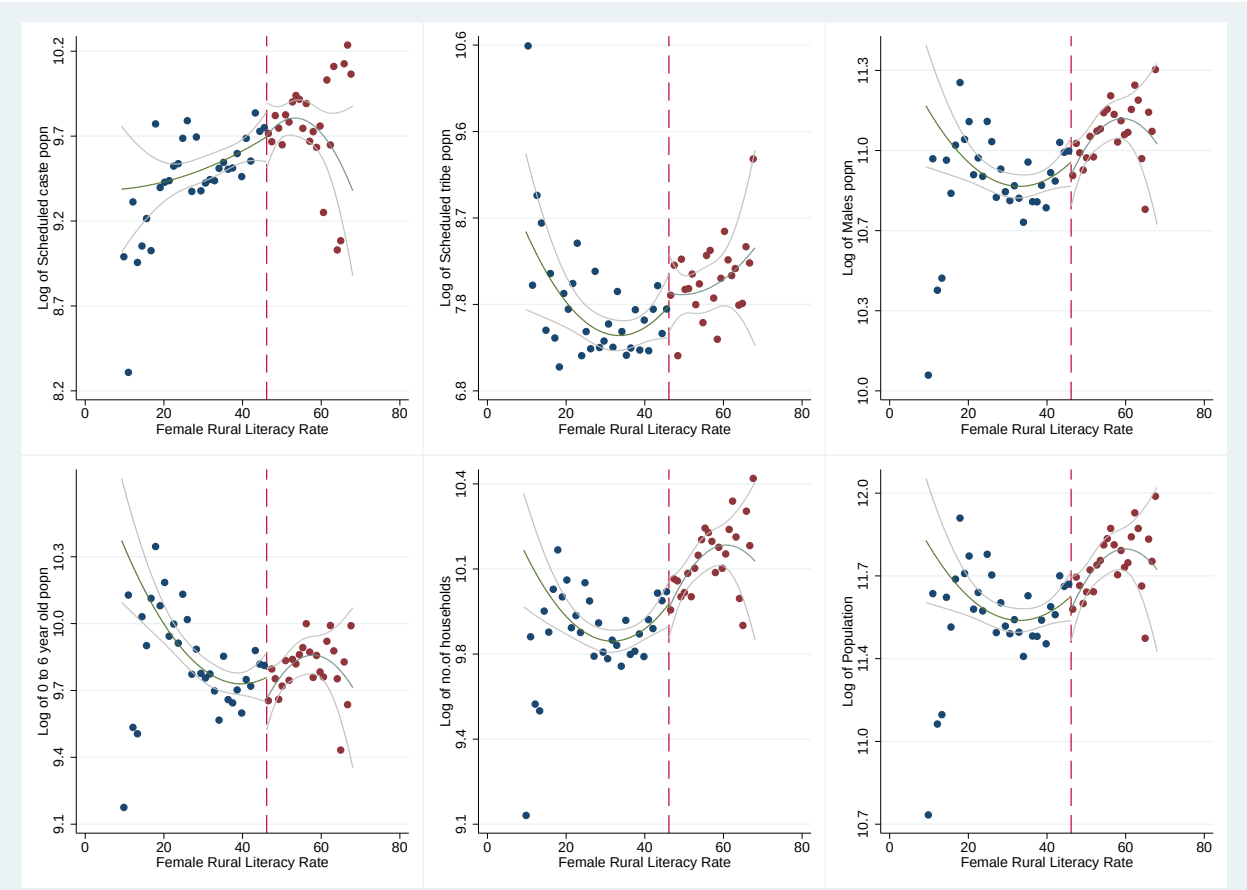
The figure plots the reduced form outcomes when the running variable is taken as the gender gap in total literacy rates for Census 2001. The sample considered satisfies the other criteria of having rural female literacy rates below the cutoff of 46.13%. We see that there is no significant discontinuity at the cutoff in either of the outcomes.

Figure B.4: Testing for Manipulation in the assignment variables



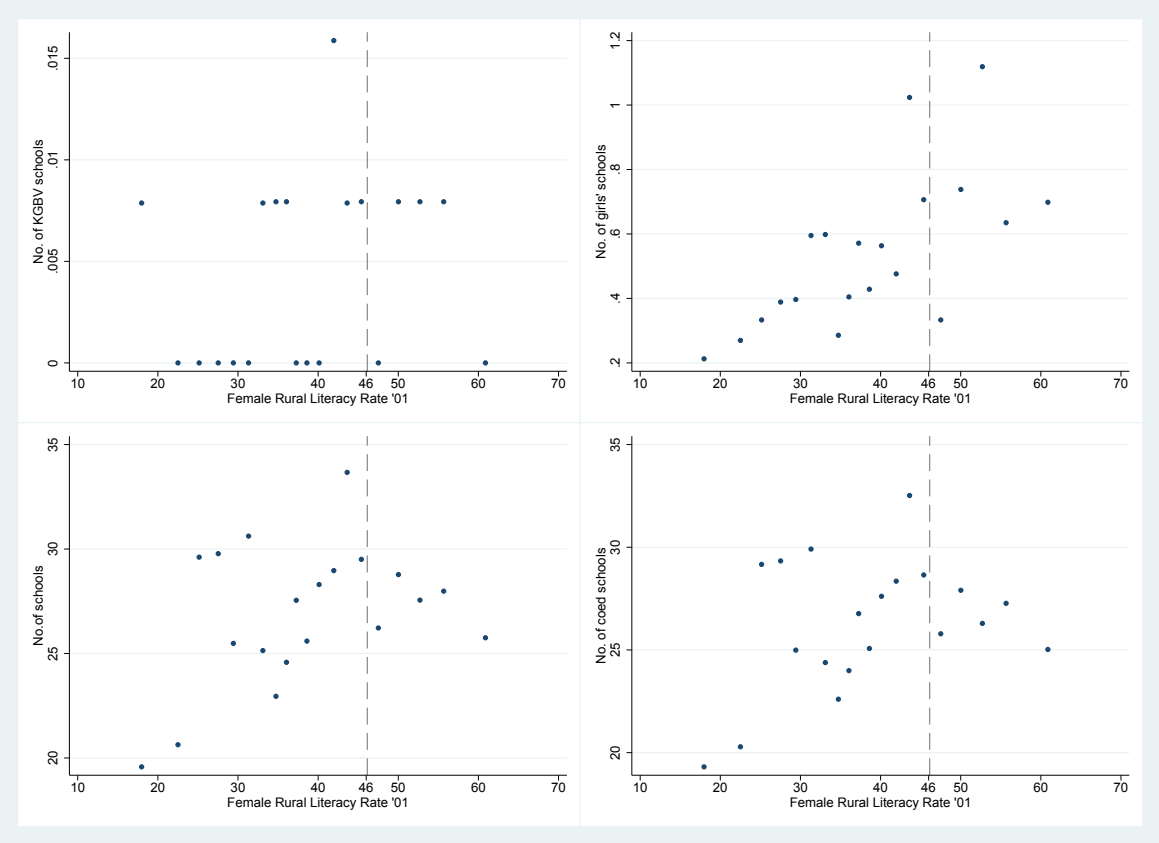
The figure shows that there was no discontinuity at the cutoff ruling out chances of manipulation.

Figure B.5: Covariates Balanced Test



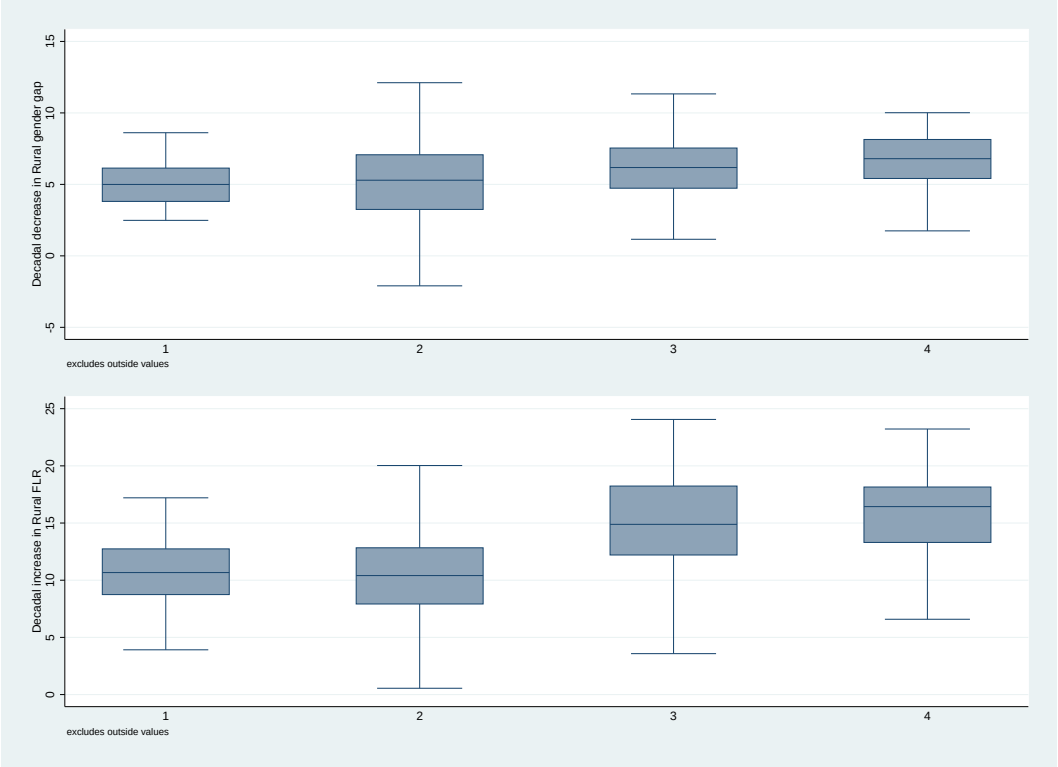
The figure shows that there was no discontinuity at the cutoff for any of the covariates.

Figure B.6: Schools built in the decade 1990-2000



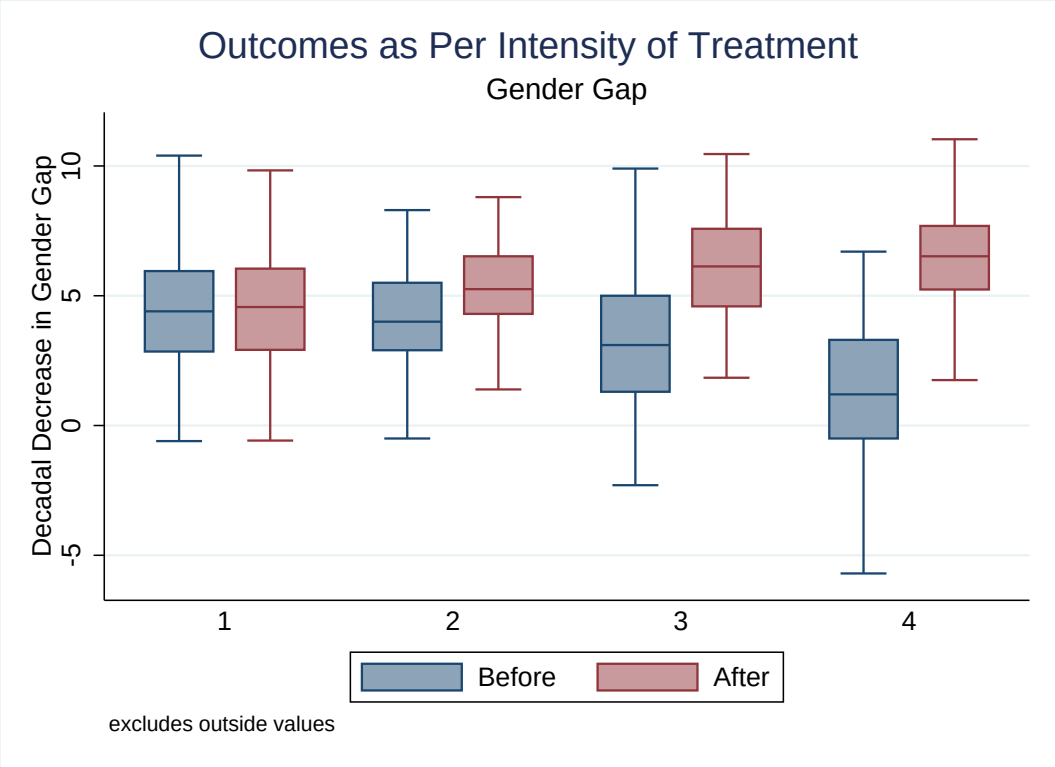
The figure plots the total number of schools built in an EBB/NEBB in the decade of 1990-2000. We see that lesser number of schools were established in the EBBs compared to NEBBs and there is no discontinuity at the cutoff.

Figure B.7: Treatment Effects at the district level



The figure shows that change in the outcome variables, increase in female literacy rate and decrease in gender gap in literacy rate in the rural areas was larger for the higher quartiles.

Figure B.8: Pre and Post treatment effects at the district level



The figure shows that the decrease in gender gap in literacy rate was significantly higher for the districts with higher concentration of EBBs

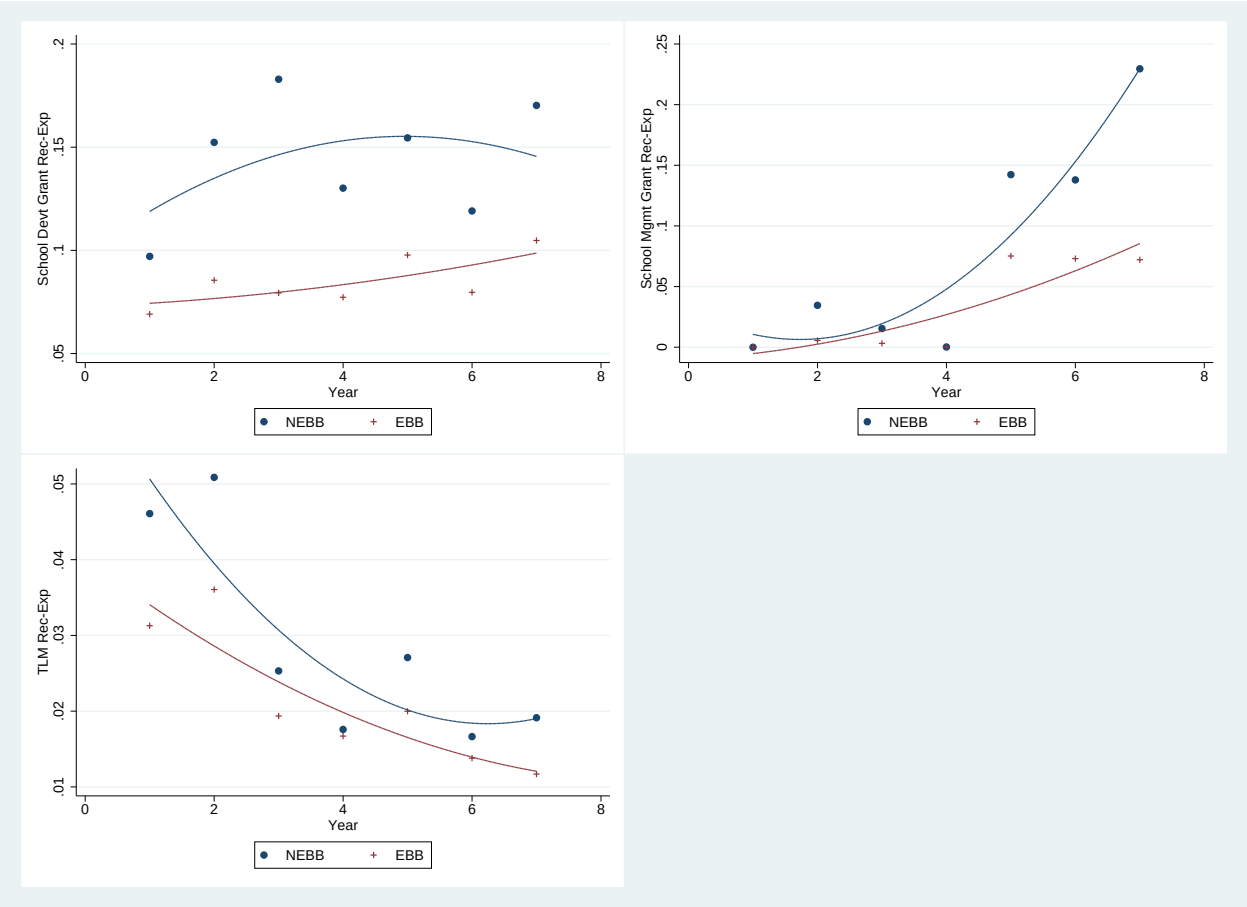
B.1.1 Analysis of Funds

The SSA funds are disbursed in the form of three grants; school development grant (SDG), school maintenance grant (SMG) and teaching learning material grant (TLM). Data on funds under NPEGEL and KGBV scheme is currently unavailable at the block level. For comparison purposes I have calculated the funds received by all the schools in a block per thousand children in thousand dollars. For the population of children I use the data from Census 2001 which reports population of children aged between 0 to 6 in rural region of the respective block. Population of children in the age group 0 to 6 may not be the best measure, but I use this as a proxy for population of children that would go to school. I also convert the value of funds measured in Indian rupees by the exchange rate for the respective academic year in which the fund was received by the school as reported and captured in the DISE data.

To have some evidence regarding the utilization of funds by schools, I use the DISE data to plot Figure B.10. The DISE data reports each of these funds received and spent by schools over the academic years 2005-2011. I aggregate this information at the block level to obtain the balance of underutilized fund by all the schools in a block. From Figure B.10 we can see the difference is positive, i.e the expenditure has always been lower than the amount received suggesting unspent balance. We can see there is more unspent balance in the NEBB and the unspent balance of the SMG has grown dramatically over the years. It is also evident that a school in EBB has received less school development grant (SDG) on average but the number has grown over the years.

The issue in doing a comprehensive analysis of fund utilization is that there is unavailability of other data on funds at the block level. The data from DISE was the only source of some information at the school level, but we do not know the components of the expenditure for diffrent facilities. Thus, using the allocations and expenditure report at state level, I find the major component of expenditure were civil works and teacher salaries. Another drawback is that SDG is received and reported only by an existing school and not by schools that are in the process of construction. I currently do not have any data to illustrate the nature and amount of funds received by a new school. Finally, the allocation of funds from the Center was at least based on some approximate methodology, however the exact rule is unclear. But, the allocation of funds to lower administrative levels seems to be based on the need of the area and schools. The exact methodology remains to be investigated.

Figure B.9: Difference in the received and expenditure amounts of the various components of SSA funds



Rec-received, Exp-Expenditure. The figure plots the difference in received and expenditure amounts of the school development grant (SDG), school maintenance grant (SMG), teaching learning material grant (TLM) over the academic years 2005-2011. The variables are measured in 1000 dollars and per thousand children in the rural region of the block. The years are labelled from 1 to 7. The graphs provide a comparison between the EBB vis-a-vis NEBB. It shows that the expenditure of funds was lower than the funds received and there is more unspent balance in the NEBB. The unspent balance of the SMG has grown dramatically over the years.

B.2 Data Appendix

India currently has 29 states, 7 union territories, 640 districts (Census 2011), but the boundaries of geographical regions have not remained the same over the decades, let alone their names. However, the boundaries of blocks, which are smaller than districts, did not change considerably in the last decade. The major hurdle in linking datasets is that there is no common ID across the data sets, hence the datasets have to be matched using names of the regions. Even the Census data for 2001 and 2011 do not have the same location code for blocks. To overcome this and match the datasets correctly, I matched data sets using the names of state-district-block. The challenge in mtaching datasets using names is that there has been change in names of the administrative divisions across years, or they have different spellings across datasets. The reason for being spelled differently mostly is because the names of administrative divisions are in the regional language, hence often there is more than one way to spell it in English.

First, I matched the Census dataset for 2001 and 2011. I use the primary census abstract tables of the census data for years 2001 and 2011 which has information on population of a region and number of literates at different administrative levels. The tables report values for the demographic variables separately for the rural, urban or total areas in a block. Based on this, the rural and total variables are created separately. For example, the rural female literacy rate of 2001 is calculated by taking the number of literate population in the rural area of the block above age 6 as a percentage of the total rural population in the block in 2001. The total literacy rate is calculated similarly but by taking the total literate population above age 6 as a percentage of the total population in the block. The number of blocks in 2001 was 5,463. In situations when I was unable to match due to change in names or spellings, I followed the strategy of renaming all the divisions according to their names in 2001 and created a unique dataset of the Census demographic variables for the years 2001 and 2011 for blocks. This however reduced my sample to 5,299 blocks for the data considering total area of the block and 5,225 blocks on the data considering rural area of the block.

Second, I match the census data with the list of classification of blocks as EBB/NEBB. I use the data for the classification from the SSA program website. The major difficulty in matching and drop in sample size happened at this stage. This was because the list of classification of blocks was prepared by the Department of Education and Literacy, and the names in this dataset were very different from those in the census. Thus, I once again followed the strategy of matching these names with the names of the census dataset by renaming the areas consistently with the Census 2001 names whenever possible. This led to the final sample size of 4,218 unique areas which is a combination of state, district and block.

The final step was to match the school census from DISE which is the unique source of information on schools in India capturing their enrollment, facilities and location. Linking the DISE data once again required to do matching based on names of the areas, as the DISE data has no corresponding code to link with the census data. This reduced my sample size further to 3,991 which is the final dataset I use throughout the paper. The drop in observations during matching should not affect the results as the drop was uniform over the classification as EBB, and hence the proportion of EBBs and NEBBs in the reduced sample was similar to the proportion in the original sample.