Candidate Selection by Parties: Crime and Politics in India

Arvind Magesan
Economics Department, University of Calgary

Andrea Szabó
Economics Department, University of Houston

Gergely Ujhelyi
Economics Department, University of Houston

June 16, 2023

Abstract

We study how parties choose candidates, a key issue to understand political selection and ultimately policy choices. Do parties select candidates that voters like, or are their choices shaped by other considerations? What is the impact of policies that limit parties' choice sets, such as limitations on candidates with a criminal history? To study these questions, we combine rich candidate level data from India with a model in which parties trade-off the electoral appeal of candidate types against internal party preferences in a strategic game of candidate selection. We find that, while parties do consider voter demand, party preferences are the dominant force in selection. Parties select criminal candidates mainly because of the *direct* payoff they yield, such as through their ability to finance their own electoral campaigns. A ban on criminal candidates can raise party payoffs by eliminating an equilibrium inefficiency. However, the ban causes voters to switch to third parties, lowering the win probability of major parties, which provides a logic for parties' unwillingness to commit to a ban on criminal candidates.

JEL: D72
Keywords: Political Parties, Elections, Representation, Voting

1 Introduction

Do parties select candidates that voters like, or are their choices shaped by other considerations? Clearly, parties like to win, and this requires running candidates that voters will support. But parties, and party leaders, may value candidate traits like loyalty, influence or wealth differently than voters do. For example wealthy candidates can self-finance their campaign, saving resources for the party. Similarly, a party may not like winning with all candidates equally. For example, a candidate further from the party's mainstream may implement more extreme policies and hurt the party's reputation - a common concern among Republicans during the 2016 electoral campaign of Donald Trump.

In a representative democracy, elections aggregate voters' preferences over candidates *who appear on the ballot*. Therefore the question of whom parties choose to run is key to understand political selection, and ultimately policy choices. A sizeable literature now explores individuals' own decision to enter politics (Besley (2005); Dal Bó and Finan (2018)). By comparison, we know much less about how parties select their candidates. As Dal Bó and Finan (2018) (p566) write: "political parties likely play a major role in who becomes a politician, and yet we have a very limited understanding of how political parties recruit and screen their candidates."

This is an especially important question in settings where there are few rules limiting whom a party can select to run as a candidate. One phenomenon of concern has been the selection of criminal candidates. In India, the widespread participation (and success) of politicians with a criminal history is well-documented and has spurred reform proposals (Vaishnav 2017).[1] Criminality in politics is also a salient issue in Brazil, where 40% of candidates running for governor in 2014 had pending court cases against them (Paiva, Sakai and Schoenster 2014). In the same year, the country introduced the Clean Record Law, banning convicted criminals from running for a period of 8 years. In the US, states vary in whether they allow convicted criminals to run for office. In Louisiana and Maine, convicted felons are eligible to run; in Massachusetts, they are ineligible to run while incarcerated, while in Texas they are ineligible even after the completion of their sentence.[2]

In this paper, we quantify some of the factors that guide parties' candidate selection strategies and analyze their implications. Studying candidate selection in Indian national elections, we combine a discrete-choice model of voter preferences over candidates with an incomplete information game of candidate selection between the two main party alliances.

[1] In 2019, 43% of candidates elected to the national legislature (the *Lok Sabha*) had faced criminal charges (https://www.hindustantimes.com/india-news/mps-with-criminal-cases-increased-in-last-decade-report-101628621064962.html).

[2] https://restoration.ccresourcecenter.org/

We use our framework to model the impact of policies that affect the set of candidates available to parties, in particular the banning of candidates with criminal backgrounds.

We begin by estimating a comprehensive BLP discrete-choice demand system describing voter preferences among candidates in Indian national elections (Berry, Levinsohn and Pakes, 1995). Voters have preferences over specific candidate characteristics, such as education, wealth, whether the candidate is Muslim,[3] and whether the candidate has a criminal background. These preferences are also shaped by constituency characteristics and various unobserved candidate characteristics.

Our specification of the supply side focuses on candidate choices by the two main party alliances in Indian national politics, the NDA and the UPA, led by the two largest parties, BJP and the INC, respectively. We use a simultaneous game of incomplete information to model the strategic interaction between these two players, and estimate their payoff functions using the Nested Pseudo-Likelihood (NPL) procedure of Aguirregabiria and Mira (2007).

Since our model is a discrete game of candidate selection, we need to define distinct types of candidates based on their observed characteristics. We approximate the pool of potential candidates that parties select from in the national election with the set of candidates contesting elections to Indian state legislatures.[4] We adopt a machine learning algorithm to identify the set of candidate clusters, or "types," that parties may consider when deciding who to run - essentially, these are combinations of candidate characteristics that tend to occur together in the data.[5] Our algorithm identifies four types: an "educated type" (educated non-Muslim with no criminal history), an "uneducated type" (uneducated non-Muslim with no criminal history), a "Muslim type", and a "criminal type" (non-Muslim with a criminal history who is also relatively wealthy).

Based on our demand estimates, we construct counterfactual vote shares to simulate all combinations of candidate types chosen by the competing parties. This allows us to construct expected vote shares and win probabilities. Parties' objective functions nest these win probabilities allowing for type-specific benefits from winning, as well as a set of heterogeneous

[3]Throughout we use Muslim to describe ethnicity rather than religion and, as we explain below, we will use individuals' names to measure this characteristic. Being Muslim (in this sense) is a particularly salient characteristic in Indian politics.

[4]A large fraction (almost 25% in our sample years) of candidates for state elections are independent candidates - many of these wanted to run for a major party but were not selected. In addition, many national level candidates begin their political careers at more local levels of politics (Dar 2019) further supporting the idea that candidates for state election approximate the pool from which national candidates are drawn.

[5]Our approach here is similar in spirit to Bandiera et al. (2020) who reduce high dimensional data on CEO activities to a small set of CEO "types" in order to study how CEOs behavior affects firm performance. Similarly, Hamilton et al. (2021) use a clustering approach to reduce the choice set of patients choosing between different medial treatments.

costs of running different candidates.

We estimate our model using detailed election data from the Election Commission of India for election years 2009 and 2014. Apart from vote shares and the number of eligible voters, the dataset also contains several candidate characteristics, such as gender and caste. We include information on candidate wealth, education, and criminal backgrounds from affidavits that candidates are required to file with the Election Commission.

Muslims are a particularly salient group in Indian politics, but this characteristic is rarely used in academic research due to a lack of data. We create an indicator for Muslim candidates based on their names using methods from text analysis. Specifically, we assign candidates to different groups based on the "distance" of their name from libraries of Muslim and non-Muslim names. This allows us to incorporate one of the most important characteristics of Indian candidates in our analysis.

Our analysis also uses information on candidates running in state elections, to create measures of both the pool of candidates that parties can choose from, and the costs faced by parties when recruiting different types from this pool. We use state elections held around the national elections we study (between 2008-2017), and assemble the same data on candidate characteristics as we do for the national elections. India's state elections are notorious for the sheer number of candidates who contest them,[6] which provides us with rich variation in candidate characteristics.

Finally, we match to our data constituency characteristics from the Indian Census. This village level information comes from the SHRUG database (Asher et al. 2020) and we use it to capture heterogeneity in voter preferences.

Our estimates of voters preferences indicate that on average voters have a particular preference for criminal types. This is especially the case among less educated and rural voters. This is in line with Vaishnav (2017), who argues that voters value criminal politicians' ability to "navigate the system" and deliver services to their constituency.

We find that considerations other than voter preferences (and hence win probabilities) also matter in parties' objective functions. All else equal, candidate types that are more common in the relevant local candidate pool are less costly for parties to recruit and run. Parties also have direct preferences over candidate types: for example both parties obtain a positive direct payoff from running criminal types relative to others. In line with Vaishnav (2017), this could reflect benefits to the party from coopting candidates with large (organized crime) networks and an ability to finance their own campaigns. In addition, parties

[6] The Indian Electoral Commission had to sharply increase the deposit that candidates pay to contest elections due primarily to the number of candidates contesting state elections, as the cost of administering these elections was becoming prohibitive in some constituencies. See Kapoor and Magesan (2018) for details.

obtain different payoffs from different candidate types conditional on winning. In particular, although voters have a strong preference for criminal candidates, parties are averse to winning with this type of candidate. This could reflect dynamic considerations of the kind discussed by Vaishnav (2016): once elected, criminal politicians may face perverse incentives, or promote some groups at the expense of others in a way that is detrimental to the parties' interests.

These payoff parameters have interesting equilibrium implications. According to our results, parties are often compelled to run a criminal candidate only because the other party is running a criminal as well. Intuitively, this happens when, faced with an opposition criminal candidate, the party's only hope for a win is to also select a criminal candidate. From the parties' joint perspective, running criminals may thus be inefficient.

Motivated by this last observation as well as recent policy proposals in India and around the world, we use our framework to model the impact of a ban on criminal candidates. One implication of the ban is to change the distribution of candidate types contesting elections, leading to higher fractions of educated, uneducated, and Muslim types.

According to our results, the ban also has implications for parties' expected winning probabilities. We find that, with a criminal ban in effect, the vote share of third party candidates rises, lowering the winning probability of both major parties. It appears that, in many cases, voters' preference for the major parties is conditional on these parties' ability to run criminal candidates.

2 Related literature

We build on and advance a growing literature that takes seriously parties' role in political selection. Galasso and Nannicini (2011) present a theoretical model where high quality candidates are valued by swing voters but are expensive for parties to recruit. This tradeoff leads to parties running high quality candidates in more competitive districts, which is consistent with data from Italy. Mattozzi and Merlo (2015) analyze a similar tradeoff, between the quality of candidates and the effort they are willing to exert to build the party organization. In their model, parties can sometimes increase effort by not running the candidate preferred by the voters. Besley et al. (2017) study a model where running the best candidate would jeopardize party leaders' survival, and find evidence consistent with their predictions in Sweden.

In this set of papers, the question is whether parties select the candidate with the highest "quality," measured in terms of education or residuals from a Mincerian wage regression. By contrast, our approach makes it possible to study selection on multiple dimensions simultaneously. We do not take an a priori stance on what constitutes a high quality candidate, and we also let the data tell us what the relevant types of candidates are in parties' choice sets.

Like us, Dal Bó et al. (2017) also considers candidate selection on many dimensions in their comprehensive study of Swedish politicians. While the focus is on self-selection, they also show evidence that parties are more willing to promote individuals who are competent (as measured by cognitive and leadership abilities) independently of their background.[7]

In developing country settings, two recent papers use field experiments that create changes in parties' nomination procedures to study the within-party selection process and why party selections may deviate from the preferred candidate of party supporters. Gulzar, Hai and Paudel (2021) provide party leaders in Nepal with information sheets on potential candidates, including information on their party service, competence, and popularity among voters. Casey, Kamara and Meriggi (2021) survey voters in Sierra Leone about their preferred candidate, and then randomize an intervention where this information is shared with party leaders, and potential candidates present their qualifications and debate each-other in a public forum. Both papers find that the interventions resulted in parties' fielding candidates closer to voters' preferences, which is consistent with a lack of information about voter preferences at baseline.

Our approach is complementary to this line of work in the sense that, although we ask a similar set of questions, we interpret observed patterns of candidate selection through the lens of a formal model, which allows for strategic interaction between parties, rather than by creating experimental changes. We also focus on different sources of preference divergence between voters and parties (e.g., regarding criminal politicians), which the literature suggests to be more relevant in the case of India.

Further, while the field experiments in Sierra Leone and Nepal yield new and important results, they also raise a number of questions. First, both experiments were conducted in relatively new democracies, where information asymmetries between parties and voters may be more severe. In our context, the well-established party organizations and a relatively long democratic history reduce the likelihood that information about voter preferences is a major constraint. Second, neither experiment was designed to study parties' strategic behavior in the candidate selection process,[8] something we explicitly account for. Third, while each study credibly identifies the average treatment effect of the intervention, neither of them was designed to identify deeper parameters, such as the weights of different objectives in parties'

[7]In Sweden socioeconomic and ethnic background likely play a considerably smaller role in politics than in India or in developing democracies more generally.

[8]In the Casey, Kamara and Meriggi (2021) experiment parties were able to choose which districts were included in the experiment, and they mostly chose safe districts with little competition.

candidate selection procedures. Our primary aim, by contrast, is to recover the deep parameters that govern party incentives and constraints, which allows answering counterfactual policy questions. Fourth, creating experimental changes in real-world nomination procedures must necessarily be limited due to ethical considerations (see Appendix 3 of Gulzar, Hai and Paudel (2021) for a detailed discussion), which creates issues for the replicability of these studies in other time periods or other countries. Our study does not face this limitation, as our methodological approach can easily be applied to other contexts.

In terms of methods, the paper closest to ours is Iaryczower, Kim and Montero (2023), who study parties' choice of policy positions (ideology) in Brazil. Like us, they combine a discrete choice specification of voter preferences over candidates with a game between parties. In their framework, candidates are exogenously given, and parties are constrained by candidates' ideology (choosing a party ideology different from candidates' ideology is costly). Our paper takes the next logical step, by studying how parties choose their candidates. In doing so we take party ideology as exogenously given, which is appropriate in a setting where the central party organization sets the policy platform and the elected candidates follow it, as seems to be the case for the two major Indian parties we study.

3 Background and data

3.1 Background

We study general elections to India's national legislature (the Lok Sabha). India constitutes a near ideal setting for studying strategic candidate selection by political parties. It is comprised of a large number of single member districts (called "constituencies"), so that in each election, voters in a given constituency elect one and only one representative from the available choices on the ballot. In contrast to a setting with proportional representation, competing parties select a single candidate to run in the constituency knowing that only voters from that constituency will be voting for him or her.

There are two main competing (pre-election) alliances, the United Progressive Alliance (UPA) and the National Democratic Alliance (NDA), led by the two main national parties, the INC and the BJP, respectively. These alliances run candidates in almost all constituencies in each election and together win the majority of seats. In every constituency, the alliance contains a group of parties (and occasionally some independent candidates) that enter into a pre-election agreement about which candidate will run to represent the alliance, without competition from other members of the group. Because our model will treat alliances as the players, we will refer to the two alliances throughout simply as *parties*.

Although we study parties' choices in national elections, we will also make use of data from state elections (elections to states' legislative assemblies). These are separate elections, and in most cases are held in different years from the national election. The constituencies in the two elections are different, in particular each national constituency is subdivided into several state constituencies. The set of parties competing in national and state elections can also be different (state elections have many regional parties), but the UPA and the NDA are major forces in state elections as well.

Indian parties are famous for their centralized organizations in which a central committee, or in some cases a charismatic leader, dictates all major decisions, including candidate selection. Farooqui and Sridharan (2014) review nomination procedures used in different countries, noting that "the USA represents the decentralised extreme, that of party primaries" while "India lies near the other extreme in that most of its major parties are at the completely or near-completely top-down of the six types of party nomination processes, with the national party leadership having the final say." (p80) Although both the INC and the BJP have formal consultation procedures that involve local party organizations in the candidate selection process, in practice decisions are ultimately made by each party's central committee (Roy (1966); Farooqui and Sridharan (2014)).[9]

Apart from electability, important factors in the candidate selection process include loyalty to the party leadership and service to the party organization. These considerations are often explicit in parties' written procedures on candidate selection (Roy 1966). Another important factor is financial considerations: financial contributions to the party and a candidate's ability to finance their own campaign. Farooqui and Sridharan (2014) (p87) describe, in the case of the BSP party, the process through which candidates are effectively bidding to receive the nomination. Similarly, Vaishnav (2017) argues that the main appeal of criminal politicians to Indian parties stems from the fact that these individuals can finance their own campaigns, including by breaking campaign finance laws if necessary.

While criminal candidates can be financially and electorally attractive to political parties, over the last several decades a host of commissions tasked with electoral reform in India have recommended the banning of criminal candidates from elections.[10] Perhaps more surprisingly, the two main parties *themselves* have at different times expressed a desire to ban criminal candidates. Yet, the parties continue to recruit criminal candidates and in July

[9] In the case of the BJP, this represents a change relative to the early 2000s, where decisions were more decentralized (Farooqui and Sridharan 2014).

[10] The Election Commission of India - Proposed Electoral Reforms of 2004 explicitly states: "The Commission is of the view that keeping a person, who is accused of serious criminal charges and where the Court is *prima facie* satisfied about his involvement in the crime and consequently framed charges, out of electoral arena would be a reasonable restriction in greater public interests. " **?**

2013, when the Supreme Court of India ruled that a sitting politician that is convicted of a criminal act should be removed from office (Lily Thomas v. Union of India) , the Government of India (with widespread support from other parties) moved to nullify the judgement.[11]

3.2 Data

We study candidate selection in India's 2009 and 2014 national elections using a dataset of official election returns combined with candidate and constituency characteristics. Election returns come from the Election Commission of India (ECI) and for each constituency they contain turnout, each candidate's name, party, and number of votes. Constituency boundaries were set in April 2008 and are unchanged throughout our sample period sample.

Our specification requires information on local (state legislative) elections matched to "corresponding" national elections. Each national election constituency contains a subset of the state election constituencies (this assignment is also constant after April 2008). In most states, state elections are held in different years from the national election, every 5 years. For each state, we assign the first state election held after 2008 to the 2009 national election and the second state election to the 2014 national election. In practice this means that state elections held between 2008-2012 are assigned to the 2009 national election and state elections held between 2013-2017 are assigned to the 2014 national election.

3.2.1 Candidate characteristics

The ECI data contains information on candidates' gender, age, and caste (Scheduled Caste, Scheduled Tribe, or General). Indian candidates for national and state elections are required to disclose their wealth and criminal history, and this information is digitized and published by the civil group ADR at www.myneta.info.[12] The ADR data also contains information on candidates' education, and we merge all this information with vote returns for both national and state elections.

Although this dataset on candidates is already quite rich, it does not contain information on one of the most important characteristics in Indian politics: whether the candidate is an ethnic Muslim. We construct a Muslim indicator from scratch based on observed candidate names as well as common fragments (substrings) of names. Specifically, we proceed as follows. First, together with the aid of a research assistant from India, we build a library of Muslim candidate names and common substrings, or "name fragments," using actual candidate names in the elections data. Muslim names are quite distinctive from other Indian

[11](`https://www.reuters.com/article/us-india-politicians-idUSBRE98N10320130924`).

[12]Previous studies using this dataset include Prakash, Rockmore and Uppal (2019), Ujhelyi, Chatterjee and Szabó (2021).

names and contain fragments that clearly distinguish them. Conversely, non-Muslim fragments would also be easy to isolate. For example "kumar" would not show up in a Muslim name and "ali" would not show up as part of a non-Muslim name. There are a total of 470 names and fragments in the Muslim library. We do the same for non-Muslim names.[13] There are a total of 1210 names and fragments in the non-Muslim library. We then compute two measures for every single name in our data which we wish to classify as Muslim or non-Muslim. The first measure is the "distance" between the name to each of the two libraries. Denote the library of Muslim names as M and the library of non-Muslim names as H. Then, for every candidate name $name_i$ in our data, we calculate the Levenshtein distance to every item in each of M and H.[14] We take the distance of $name_i$ and library M to be the minimum of these distances for names in library M. Similarly, the distance between $name_i$ and library H is the smallest distance between $name_i$ and all names in H. Let $d(name_i, M)$ and $d(name_i, H)$ denote these distances.

Next, we use the name fragments to construct another measure. Specifically count, for every name in the data, how many Muslim fragments and how many non-Muslim fragments appear in the name, and divide this by the number of fragments in the respective library to get a frequency. Denote these as $frag(name_i, M)$ and $frag(name_i, H)$ respectively.

Finally, $name_i$ is assigned "Muslim" identity if either $frag(name_i, M) > frag(name_i, H)$ or $\{frag(name_i, M) = frag(name_i, H)$ and $d(name_i, M) < d(name_i, H)\}$, and it is assigned "non-Muslim" identity otherwise.

3.2.2 Constituency characteristics

In the Indian electoral system, some constituencies can only be contested by Scheduled Caste or Scheduled Tribe candidates and the ECI data contains indicators for these reserved constituencies.

For demographic and other characteristics of each constituency, we use the SHRUG dataset (Asher et al. 2020). Specifically, we use village-level information from the 2011 Indian Census, which the SHRUG allows to be matched to constituencies. We use the following characteristics: literacy rate, share of working population, share of Scheduled Caste and Scheduled Tribe population, whether the village has access to paved roads, and whether the village is located in a rural or urban area. We use both the village level information, and also aggregate it up to the constituency level.

[13] While the majority of non-Muslim names will have Hindu or Sikh origins, there is also a substantial population with Christian names.

[14] The Levenshtein distance simply counts the number of single edits required to turn one string into another. For example the Levenshtein distance between "car" and "stare" is 3.

Our dataset is limited by the constituencies for which we can obtain demographic information. The main reason we are forced to drop constituencies is that some state-level constituencies are missing from the SHRUG because the villages they contain could not be uniquely matched to them (see Asher et al. (2020)). Because we analyze national elections, there is another step involved in aggregating the state constituencies up to the national constituency level. There are reasons to believe that state constituencies that are missing in a national constituency are systematically different from other state constituencies (e.g., they are more likely to be large urban areas). Therefore we only include in our analysis national constituencies for which we have information on all the state constituencies they contain. This drops from the sample several states in their entirety (mostly small states with only a few national constituencies).[15] From the remaining 18 states, we drop 3 because we either only have less than 20% of their constituencies or because they have very few constituencies to begin with.[16] The remaining 15 states contain 478 of the 538 constituencies in India, and we have constituency characteristics from the SHRUG for 234 of these. We drop 2 constituencies because some of their candidates have unrealistically high numbers of criminal convictions,[17] leaving us with a total of 232 constituencies in the dataset. These constituencies are contested by a total of 3208 candidates in 2009 and 3373 candidates in 2014. The 232 national constituencies contain 1629 state constituencies, and these state constituencies are contested by 17,965 candidates in the 2009.election period and 18,801 in the 2014 election period. Summary statistics of our data are in Table 1.

4 Model

We consider a simple simultaneous move Bayesian game of candidate selection between competing parties. Candidates are described by a set characteristics. In selecting a candidate, each party weighs its own internal preference over candidates against the preferences of voters, and thus the probability of winning. We discuss the decision problem of parties and voters in turn.

[15] The states excluded and the total number of their national constituencies are: Arunachal Pradesh (2), Goa (2), Manipur (2), Meghalaya (2), Mizoram (1), Nagaland (1), Puducherry (1), Punjab (13), Sikkim (1), Tripura (2), Uttarakhand (5), Delhi NCT (7).

[16] Specifically, we drop Chhattisgarh, with only 2/11 constituencies, Himachal Pradesh, with 2/4 constituencies, and Jammu & Kashmir, with 1/6 constituencies.

[17] Both of these are in Tamil Nadu, and both have a candidate with close to 400 criminal cases.

Table 1: Summary statistics

	N	Mean	Std. Dev.	Median	10%	90%
A. Candidate characteristics - state elections						
UPA (0/1)	36766	0.10				
NDA (0/1)	36766	0.11				
Education (0/1)	29137	0.62				
Muslim (0/1)	36766	0.11				
Criminal history (0/1)	30465	0.20				
Assets (log)	29903	13.90	2.36	14.00	10.84	16.83
Male (0/1)	36766	0.93				
Age	36766	44.76	11.27	44	31	61
B. Candidate characteristics - national elections						
UPA (0/1)	6581	0.07				
NDA (0/1)	6581	0.07				
Education (0/1)	2900	0.75				
Muslim (0/1)	6581	0.12				
Criminal history (0/1)	3068	0.25				
Assets (log)	3012	14.78	2.52	14.97	11.55	17.71
Male (0/1)	6581	0.93				
Age	6581	46.25	11.96	45	31	63
SC or ST (0/1)	6581	0.35				
C. Constituency characteristics - national elections						
Eligible voters (1000)	464	1421.50	195.59	1426.24	1173.14	1685.34
Turnout	464	64.97	12.39	65.99	47.57	81.05
N. of candidates (before aggregation)	464	14.18	6.06	14	7	22
N. of candidates (after aggregation)	464	5.71	1.46	5	4	8
Reserved constituency (0/1)	464	0.28				
Literate population (%)	464	0.61	0.10	0.61	0.49	0.74
ST and SC population (%)	464	0.27	0.14	0.24	0.13	0.46
Rural population (%)	464	0.82	0.11	0.83	0.68	0.94
Population with paved roads (%)	464	0.83	0.19	0.90	0.58	1.00
Working population (%)	464	0.42	0.07	0.43	0.32	0.50

Notes: Education: 1 if completed high school. Criminal history: 1 if has at least one criminal case. N. candidates (after aggregation) is number of candidates once small-party candidates are aggregated into one as described in the text.

4.1 Parties

Consider an electoral constituency where competing parties choose which candidate to run for election. Each party $p \in \{NDA, UPA\}$ chooses one candidate out of a set of potential candidates $\mathcal{A}_p$, where $|\mathcal{A}_p| = K_p$. Let the choice of party p be given by a_p, and denote the vector of choices of p's opponents by $\mathbf{a}_{-p}$. Voters cast their votes based on the candidates that parties choose to run: let $w_p(a_p, \mathbf{a}_{-p})$ represent the winning probability of party p associated with a candidate selection profile $(a_p, \mathbf{a}_{-p})$.

A key innovation of our approach is to allow for the fact that parties may care about the candidate they run beyond its effect on the vote share s_p. First, a party may experience costs or benefits from a candidate it runs *directly* (i.e., independently of the vote share). This could reflect considerations such as the availability of certain types of candidates (e.g., a party with few Muslim members may find it more costly to run a Muslim candidate), internal politics (e.g., some candidates may be loyal to the party leadership, while others may come from a competing faction within the party) or party finances (e.g., some candidates may be able to finance their own campaigns, making them a "cheaper" choice for the party). Second, the choice of candidate could also *mediate* the party's payoff from a larger vote share. For example, a party may prefer winning with a traditional candidate than with an outsider who will disrupt politics and policy making.

To capture these considerations, we specify party p's payoff from choosing candidate $a_p \in \mathcal{A}_p$ as

$$b(a_p) \times w_p(a_p, \mathbf{a}_{-p}) + c(a_p) + \varepsilon_p(a_p), \tag{1}$$

where $b(a_p)$ represents the mediating effect of a_p on the party's payoff from the winning probability w_p, while $c(a_p) - \varepsilon_p(a_p)$ captures payoffs from a_p that are independent of w_p. For clarity, we will refer to these direct payoffs as "costs" (although they could be positive, i.e., a benefit). The difference between $c(a_p)$ and $\varepsilon_p(a_p)$ is that the former is observable to all competing parties, while the latter is party p's private information. The private component $\varepsilon_p(a_p)$ is distributed i.i.d. across parties and candidates, with cdf $G(.)$.

An important factor affecting $c(a_p)$ is the pool of potential candidates available to the party. This will be determined by who the party's members are, and who among its members has both the motivation and ability to run for office. As explained below, we will measure a party's pool of potential candidates using information on the party's candidates in local elections. This is motivated by the fact that (i) many Indian parties competing in local elections have clear affiliations to a national party (either the party is the same, or they belong to the same electoral alliance), and (ii) it is common for national politicians to begin their political careers in local elections. To highlight this, write $c(a_p) = c(a_p, L_p)$, where L_p

denotes the pool of party p's candidates in relevant local elections.

Given the presence of private information, this setup gives rise to a simultaneous game of incomplete information between parties competing in the constituency. The solution concept is Bayesian Nash Equilibrium (BNE). For a realization of the private costs $\varepsilon_p \equiv \{\varepsilon_p(a)\}_{a \in \mathcal{A}_p}$, a party chooses candidate $a_p(\varepsilon_p)$. Let $P(\mathbf{a})$ denote the ex ante probability of a profile of choices $\mathbf{a}$. Then given ε_p and $P(\mathbf{a}_{-p})$, in a BNE party p chooses a_p to maximize its expected payoff given all other parties' strategies:

$$a_p \in \arg\max_a U_p(a, P(\mathbf{a}_{-p})),$$

where

$$U_p(a, P(\mathbf{a}_{-p})) \equiv b(a) \times E_P[w_p(a, \mathbf{a}_{-p})|a] + c(a, L_p) + \varepsilon_p(a) \tag{2}$$

is the expected value of (1) over the possible realizations of opponents' choices $\mathbf{a}_{-p}$.

For the purposes of estimation it is convenient to express strategies as *choice probabilities* (CPs). In particular, define payoffs net of the private cost as

$$\tilde{U}_p(a, P(\mathbf{a}_{-p})) \equiv U_p(a, P(\mathbf{a}_{-p})) - \varepsilon_p(a) \tag{3}$$

so that a_p maximizes party p's expected payoffs iff:

$$\tilde{U}_p(a_p, P(\mathbf{a}_{-p})) + \varepsilon_p(a_p) \geq \tilde{U}_p(a, P(\mathbf{a}_{-p})) + \varepsilon_p(a) \quad \forall a \in \mathcal{A}_p.$$

The probability of party p choosing action a_p given the opponent's strategy $P(\mathbf{a}_{-p})$ is then:

$$\begin{aligned} P(a_p) &= \int_{\varepsilon_p} \mathbf{1}\Big\{\varepsilon_p(a) - \varepsilon_p(a_p) \leq \tilde{U}_p(a_p, P(\mathbf{a}_{-p})) - \tilde{U}_p(a, P(\mathbf{a}_{-p})), \quad \forall a \in \mathcal{A}_p\Big\} dG(\varepsilon_p) \\ &\equiv \Lambda_p(a_p; \mathbf{P}_{-p}) \end{aligned} \tag{4}$$

Equilibrium in the game is fully characterized by a fixed point in $P(\mathbf{a})$ of the system of equations defined by (4) for all p. Stacking equations by actions and parties, an equilibrium vector of CPs $\mathbf{P}^*$ satisfies:

$$\mathbf{P}^* = \mathbf{\Lambda}(\mathbf{P}^*)$$

Under the assumption that $\varepsilon_p(a)$ follows the Type 1 Extreme Value Distribution, the equilibrium CPs satisfy

$$\Lambda_p(a_p; \mathbf{P}_{-p}) = \frac{\exp\left\{\tilde{U}_p(a_p, \mathbf{P}(\mathbf{a}_{-p}))\right\}}{\sum_a \exp\left\{\tilde{U}_p(a, \mathbf{P}(\mathbf{a}_{-p}))\right\}} \ \forall p.$$

4.2 Voters

To model parties' winning probabilities w_p as a function of the set of candidates running, we consider the individual decisions made by a continuum of voters. We assume expressive voting with a flexible specification of voter preferences over candidates' characteristics (Ujhelyi, Chatterjee and Szabó (2021) - USC (2021) from now on). [18]

Specifically, each candidate can be described by a vector of characteristics $\mathbf{x}$, such as their education level or criminal history. Given a set of candidates that parties have chosen to run, voter i's utility from voting for the candidate of party p is

$$V_{ip} = \beta_i \mathbf{x}_p + \xi_p + \eta_{ip}. \tag{5}$$

The first term represents voters' (potentially heterogenous) preferences over the characteristics of p's candidate. The second term, ξ_p, allows for unobserved (to the researcher) candidate characteristics valued by voters or, equivalently, shocks to parties' popularity in the given constituency. The distribution of ξ_p is left unspecified, and it can be correlated with $\mathbf{x}_p$. Finally, η_{ip} are individual preference shocks drawn from a Type-I Extreme Value distribution. To model the sources of preference heterogeneity among voters, write

$$\beta_i = \beta + \mathbf{\Pi} \mathbf{d}_i, \tag{6}$$

where $\mathbf{d}_i$ is a vector of voter demographics, while β and $\mathbf{\Pi}$ contain the parameters.

To complete the voter's choice set, let $p = 0$ indicate the option to abstain and

$$V_{i0} = \pi_0 \mathbf{d}_i + \eta_{i0}$$

the voter's associated utility. This allows for the utility of abstention (hence the cost of voting) to vary across voters.

Voter i chooses option p (vote for one of the parties or abstain) if $V_{ip} > V_{ip'}$ for all $p' \neq p$.

[18] The assumption of expressive voting is supported by extensive survey evidence on Indian voters' motivations. Banerjee (2017) provides a book-length discussion of the meaning that voters attach to the act of voting. Based on a recent survey, (Heath and Ziegfeld 2022) estimate that at most 1.1% of individuals vote strategically

Thus, voters choose between their options based on the observed and unobserved candidate characteristics, the benefit of abstention, and their idiosyncratic shocks. This implicitly defines the set for which voter i will choose option p, $\{(\mathbf{d}_i, \eta_i)|V_{ip} > V_{ip'} \text{ for all } p' \neq p\}$. Given a distribution of $\mathbf{d}_i$ and η_i, integrating over this set yields parties' vote shares as a function of their candidate choices. Under the assumed Type-I EV distribution for η_{ip} and given a distribution $F(\mathbf{d}_i)$, these vote shares can be written as

$$s_p(a_p, \mathbf{a}_{-p}) = \int \frac{\exp\left[\beta_i \mathbf{x}_p + \xi_p - \pi_0 \mathbf{d}_i\right]}{1 + \sum\limits_{q>0} \exp\left[\beta_i \mathbf{x}_q + \xi_q - \pi_0 \mathbf{d}_i\right]} dF\left(\mathbf{d}_i\right). \tag{7}$$

Expected winning probabilities are then given by

$$E_P[w_p(a_p, \mathbf{a}_{-p})|a_p] = \sum_{\mathbf{a}_{-p}} \mathbf{1}\left\{s_p > s_{p'} \ \forall p' \neq p\right\} P(\mathbf{a}_{-p}). \tag{8}$$

5 Specification and estimation

5.1 Overview

Our ultimate goal is to estimate parameters in the parties' objective function (2). To do this, we proceed in two stages. In the first stage, we estimate voters' utility functions (5) with a BLP procedure, using as instruments variables that enter parties objective function but do not directly enter voter utilities. This yields estimates of the voter preference parameters β and $\mathbf{\Pi}$, as well as the popularity shocks ξ. Armed with these estimates, we can use our model of voters to predict parties' vote shares given *any* combination of candidates, based on (7). In the second stage, we use these estimated vote share functions to estimate the parameters of the $b(.)$ and $c(.)$ functions in (2) using a Pseudo-Maximum-Likelihood procedure.

5.2 Estimating voter preferences

Estimation follows the Generalized Method of Moments (GMM) algorithm proposed by Berry, Levinsohn and Pakes (1995). Detailed treatments of the procedure can be found in Berry, Levinsohn and Pakes (1995),Nevo (2000) and Nevo (2001)). Here we modify a previous application of this procedure to Indian *state* elections in USC (2021)

5.2.1 Specification and endogenous characteristics

To deal with the presence of many small parties and independent candidates, we follow USC (2021) and aggregate these candidates in each constituency. Specifically, in each constituency we aggregate into one "small party" category parties that are not part of either the UPA or the NDA alliance and only run a few candidates in the data. We focus on four candidate characteristics $\mathbf{x}$: education, Muslim, crime, and assets.[19] To be consistent with the estimation of party objectives below, we standardize all these variables to have 0 mean and unit standard deviation. We also include in this vector an indicator for candidates where one or more characteristics were imputed. In addition, we include in (5) the following control variables: party and alliance fixed effects (to control for a portion of ξ_p that is common across constituencies), state and year fixed effects and an indicator for reserved constituencies (to allow for different payoffs from abstention).

The BLP procedure necessitates the use of instrumental variables (IV) for two reasons: first, to identify the "nonlinear" parameters $\mathbf{\Pi}$, and second, to identify the parameters on any variables in $\mathbf{x}$ that parties can adjust in response to the popularity shocks ξ_p (the usual endogeneity problem). Because the focus of our study is parties' choice of their candidates, we treat all four candidate characteristics as endogenous. Variables that enter parties' objective function (2) but do not directly enter voter utility are valid instruments - in our specification, these are parties' candidates in local elections, L_p. Recall the idea behind the presence of these variables in (2): a party's available pool of candidates affects its cost of choosing candidates with specific characteristics. For example, a party has a lower cost of finding a highly educated candidate when most candidates in its pool are highly educated. The prevalence of high education (for example) in the pool of candidates is measured by the prevalence of this characteristic among the candidates a party runs in local elections in the same geographic area. These variables are valid instruments as long as voter valuations ξ_p in the national election are uncorrelated with the characteristics of a party's candidates in the local election.[20]

We create our instruments based on party alliance (UPA/NDA/neither) and the state assembly constituencies overlapping with the national election constituency. For example, we instrument the assets of an UPA candidate with the average assets of all UPA candidates running in the assembly constituencies contained in the given national election constituency. We create these instruments both for the same election and for the other election in the data.

To get a preliminary sense of the relevance and strength of these instruments (which

[19] We also considered three other characteristics observed in the data, caste, gender, and age. Gender has very little variation (almost all candidates are male). Caste has very little variation once we control for constituency reservation. Age does not seem to be an important characteristic for either voters or parties in these elections.

[20] The idea corresponds to Industrial Organization applications that use cost shifters that affect firm profits but not consumer utilities as instruments for endogenous variables (typically, prices). USC (2021), which studied state elections, used candidate characteristics in neighboring constituencies based on a similar logic.

we also explore in more detail below), we regress each characteristic on the corresponding instruments and control variables. The results are shown in Table 2. In columns (1) and (2), the state election averages of the education characteristic have large and significant association with the education level of a party's candidates in the national election. Columns (3-8) show similar patterns for the other characteristics as well. These correlations, which are interesting in their own right, provide support for the idea that state candidate averages proxy for the pool of characteristics that a party is able to draw from at the national level.

Table 2: Characteristics regressed on instruments

Dep. var.:	Education		Muslim		Crime		Assets	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mean dep. var.	0.14	0.14	-0.03	-0.03	0.05	0.05	-0.10	-0.10
Std. dev.	0.84	0.84	0.88	0.88	0.96	0.96	1.00	1.00
IV1	0.19	0.16	0.54	0.34	0.29	0.28	0.32	0.31
	(0.05)	(0.05)	(0.07)	(0.08)	(0.06)	(0.06)	(0.05)	(0.05)
IV2		0.34		0.44		0.10		0.10
		(0.08)		(0.10)		(0.09)		(0.08)
Adj. R^2	0.10	0.11	0.09	0.09	0.14	0.14	0.37	0.37
F	13.23	15.22	60.73	42.73	22.94	13.34	42.73	20.90

Notes: IV1 is the average of the given characteristic among an alliance's candidates in state election constituencies in the given year. IV2 is the same variable for the other election in the data (2009 for 2014 and vice versa). Regressions control for state, year, party and alliance fixed effects, and indicators for imputed characteristics and reserved constituencies. The F statistic is the Olea and Pflueger (2013) effective F statistic. Robust standard errors in parentheses. N = 2649.

5.2.2 Nonlinear parameters and differentiation IVs

In order to identify nonlinear parameters, we follow recent work by Gandhi and Houde (2019) and use the instruments just described to create "differentiation IVs." The idea behind these instruments is to use the menu of choices available to each decision-maker to identify preference heterogeneity among decision-makers. In our application, preference heterogeneity among voters for a candidate's education level (say) is identified based on how many candidates in a voter's choice set have similar education levels.

To construct the differentiation IVs, we first predict each endogenous candidate characteristic using the above instruments (and all exogenous variables). We then use these predicted characteristics to form instruments: for each (predicted) characteristic of a candidate that enters the nonlinear part of voter utility, we compute the number of candidates in the constituency whose corresponding (predicted) characteristic is within one standard deviation. We also use the methods proposed by Gandhi and Houde (2019) to guide our specification choice and evaluate the strength of our instruments. The idea is to evaluate whether the instruments are "strong enough" to reject the linear (Logit) specification, i.e., $\mathbf{\Pi} = \mathbf{0}$. First, we enter the differentiation IVs as controls in a Logit specification. This specification still includes all the controls described above, and instruments the endogenous characteristics with the 8 instruments created from the state election data. The results are in Table 3 In column (1), the differentiation IVs for Muslim and assets are statistically significant while the differentiation IVs for education and crime are not. This suggests that the former two are capable of capturing departures from the Logit model. As an alternative diagnostic, we also run a specification that includes the differentiation IVs as instruments instead of controls. The last row of the table (IIA p-val) shows the p-value of the overidentification J-test for this specification. The fact that this specification is clearly rejected also provides support for focusing on the nonlinear specifications (Gandhi and Houde 2019). In column (2) we use only the differentiation IVs for Muslim and assets and obtain similar conclusions.

To further evaluate the nature of preference heterogeneity, we estimate a random coefficients specification where voter demographics $\mathbf{d}_i$ in (6) are replaced with random variables drawn from a standard normal distribution. We use a separate i.i.d. variable for each of the four candidate characteristics, and estimate this specification using the BLP procedure with the differentiation IVs as instruments. This specification, displayed in column 3 of Table 3, indicates the presence of significant heterogeneity in voters' preference for candidate assets, but not for the other three characteristics.

5.2.3 Adding voter demographics

Based on the specification checks described in the previous section, we first focus on identifying the relevant sources of preference heterogeneity in voters' valuation of candidate assets. Our main demographic variables are literacy, rural population, presence of paved roads, working population, and lower caste population. Following Gandhi and Houde (2019), we interact the differentiation IV for assets with the average value of each of these demographic variables in the constituency.[21]

We again evaluate these instruments using a Logit specification. This supports using the interaction of the asset differentiation IV with literacy, rural population, and presence of paved roads (Table 4). Estimating nonlinear specifications using different combinations of these instruments and corresponding nonlinear parameters yields a clear favorite, shown in Column (1) of Table 5. This specification passes the overidentification J test and results

[21] The idea is to identify a parameter π_m^k on the interaction of the demographic d_m with the characteristic x^k using the instrument $\bar{d}_m \hat{x}^k$, where $\bar{d}_m$ is the average value of the demographic in the constituency and $\hat{x}^k$ is the differentiation IV for the candidate characteristic.

in nonlinear coefficients that are jointly statistically significant (based on the Newey-West test).

According to the estimates in column (1), all else equal voters dislike Muslim candidates and like candidates with a criminal history (consistent with Vaishnav (2017) and USC (2021)) though the latter is not statistically significant. Rural voters have a preference for candidates with more assets. A possible explanation is that wealthier candidates have higher social status and thus easier access to other government officials to advance local interests and "get things done." According to Vaishnav (2017), providing these connections between the local community and the state apparatus is very important in voters' evaluation of the candidates.

Table 3: Specification choice: differentiation IVs and random coefficients

	Logit (1)	Logit (2)	Random coefficients (3)
education	-0.63	-0.62	-0.38
	(0.40)	(0.40)	(0.51)
Muslim	-0.37	-0.39	-0.42
	(0.17)	(0.17)	(0.32)
crime	-0.00	0.16	0.33
	(0.34)	(0.31)	(0.38)
assets	1.67	1.66	2.65
	(0.27)	(0.28)	(0.56)
diffIV(educ)	0.05		
	(0.03)		
diffIV(Muslim)	-0.07	-0.08	
	(0.03)	(0.02)	
diffIV(crime)	-0.04		
	(0.03)		
diffIV(assets)	-0.08	-0.07	
	(0.04)	(0.04)	
$\pi_{education}$			-0.04
			(8.00)
π_{Muslim}			-0.06
			(6.82)
π_{crime}			0.04
			(9.82)
π_{assets}			-1.63
			(0.46)
J p-val	0.10	0.07	0.02
IIA p-val	0.00	0.00	

Notes: Columns (1) and (2) are specification checks proposed by Gandhi and Houde (2019). The dependent variable is vote shares. Candidate characteristics are instrumented with the instruments described in section 4.2.1, and the "differentiation IVs" are entered as controls. Specifications also control for state, year, party and alliance fixed effects, indicators for imputed characteristics, and reserved constituencies. J p-val is the p-value of the overidentification J test. IIA p-val is the p-value of the overidentification J test when the diffIV variables are used as instruments instead of controls. Column (3) is a random-coefficients specification, using standard Normal draws instead of voter demographics, and using the four diffIV variables as instruments. Robust standard errors in parentheses.

Table 4: Specification choice: differentiation IVs and constituency demographics

	(1)	(2)	(3)	(4)	(5)	(6)
diffIV(assets x literacy)	-0.20					-0.30
	(0.05)					(0.16)
diffIV(assets x rural)		-0.09				0.17
		(0.04)				(0.09)
diffIV(assets x roads)			-0.15			-0.18
			(0.04)			(0.07)
diffIV(assets x workers)				-0.24		0.08
				(0.08)		(0.26)
diffIV(assets x caste)					-0.12	0.12
					(0.09)	(0.12)
J p-val	0.07	0.04	0.06	0.05	0.03	0.20
IIA p-val	0.00	0.00	0.00	0.00	0.02	0.00

Notes: Specification checks proposed by Gandhi and Houde (2019). The dependent variable is vote shares. Candidate characteristics are instrumented with the instruments described in section 4.2.1, and the "differentiation IVs" are entered as controls. Only the coefficients on the differentiation IVs are shown. Specifications also control for state, year, party and alliance fixed effects, indicators for imputed characteristics, and reserved constituencies. J p-val is the p-value of the overidentification J test. IIA p-val is the p-value of the overidentification J test when the diffIV variables are used as instruments instead of controls. Robust standard errors in parentheses.

Table 5: Voter preference parameter estimates

	(1)	(2)	(3)	(4)
Linear parameters (β)				
education	-0.19	0.94	-0.59	-0.57
	(0.69)	(0.82)	(0.73)	(0.80)
Muslim	-0.50	-0.37	-0.54	-0.54
	(0.28)	(0.27)	(0.29)	(0.30)
crime	0.66	0.73	0.56	0.59
	(0.51)	(0.57)	(0.50)	(0.50)
assets	2.40	-2.00	2.81	4.84
	(3.48)	(1.14)	(2.72)	(6.05)
Nonlinear parameters (π)				
assets x literacy	-3.45	4.15		-2.42
	(3.49)	(1.85)		(5.95)
assets x road		2.59	-3.78	-3.60
		(1.50)	(2.85)	(3.37)
assets x rural	3.84		3.97	3.26
	(1.27)		(1.21)	(2.47)
J	1.481	9.715	8.880	9.407
df	4	4	4	4
p-value	0.83	0.05	0.06	0.05
Newey-West pval	0.001	0.005	0.007	0.001

Notes: BLP estimates. J is the overidentification J-statistic with its degree of freedom (df) and p-value. Newey-West pval is the p-value of the Newey-West D-test of the null that all nonlinear parameters are jointly 0. Robust standard errors clustered by constituency in parentheses.

5.3 Specifying the parties' choice sets

In our specification of voter preferences, we conceptualized parties' choice of candidates a_p as a choice of a bundle of candidate characteristics $\mathbf{x}_p$. Applied directly to parties' problem, this would imply very large choice sets $\mathcal{A}_p$, containing all the possible combination of characteristics. This is neither practical for estimation, nor realistic as a model of party choices. For example, it is unlikely that a party would view two candidates who are identical in all dimensions but whose assets are slightly different as substantively different options.

For a better model of parties' problem, we assume the existence of a smaller set of candidate "types" that parties consider when choosing who to run. For example, a type could be an "educated non-Muslim with some criminal history in the second quartile of the asset distribution." Rather than constructing these types ourselves, we use machine learning tools to let the data tell us what they should be.

5.3.1 Data and variables for constructing candidate types

Our goal is to describe the pool of potential national candidates (as opposed to the set of candidates actually selected by the parties). As argued above, the candidates running in state elections provide a good proxy for the pool of candidates that national parties can select from. Thus, we define candidate types based on the characteristics of candidates running in state elections.

Using state candidates to define types also has practical advantages: there are many more state candidates, so we can use more observations to create the types; and we observe more complete candidate characteristics at the state level. Specifically, we have more information on independent candidates' characteristics in this data - these candidates are often individuals who wanted to run with a major party but were not selected.

The state election data has 36,766 observations. We drop 2 observations because they have missing characteristics *and* have no similar candidates (based on gender, caste and age) that we could use for imputing these missing values (see details below). As above, we use the candidate characteristics education, assets, criminal history and Muslim.

Education, assets, and criminal history have missing values, which we need to impute in order to assign each candidate to a type.[22] For imputation, we rely on the candidate's gender, age and caste, which have no missing values.

[22] An alternative approach we considered is to use a Missing indicator as an additional characteristic. However, there are differences between missing values in the state and national election data (for example, most independent candidates' education, assets, and criminal history is missing in the former). Thus, using Missing as an additional characteristic would mechanically make candidate types in the two datasets less comparable.

The asset variable has 6301 missing values and an additional 562 zeros which we also treat as missing. For missing values, we impute the average value of assets by gender, caste and age range, where the age range is specified as +/- 1 year relative to the candidate's age. For example, a 30 year old male general caste candidate's imputed asset is the average of all male general caste candidates aged 29-31. After imputation, we use $\log(assets + 1)$ as our asset characteristic.

Criminal history has 6301 missing values. We impute the number of criminal cases using the average by gender, caste and +/-1 year age bin, and then use an indicator equal to 1 if the number of cases is at least one, and 0 otherwise.

Education has 7629 missing values. We impute the number of completed years of education using the average by gender, caste and +/-1 year age bin, and then use an indicator equal to 1 if years of education is at least twelve, and 0 otherwise.

All four of our clustering variables are standardized to have 0 mean and unit standard deviation.

5.3.2 Clustering algorithm

We use k-means clustering to create the types. This iterative procedure partitions the data into K clusters based on the 4 variables described above (Muslim, education, assets and crimes). The algorithm begins by specifying K initial centroids and forming clusters by assigning each candidate to the closest centroid. Throughout, we use Eucledian distance to compute candidates' distance from a centroid. Next, new centroids are computed based on the average characteristics of the candidates assigned to each cluster. Using these centroids, candidates are reassigned to the closest cluster, and the process continues until no candidate is reassigned from their current cluster.

To use k-means clustering, one must first choose the number of clusters K. There is currently no cross-validation method to assess the relative performance of different values, and one option is to choose the number K based on substantive considerations (Athey and Imbens (2019)).[23] Alternatively, we can adopt a set of commonly used measures in the machine learning literature to select K. The first measure relies on the Within Cluster Sum

[23] Alternative unsupervised methods such as Density Based Clustering (DBSCAN) do not require the researcher to input the number of types K. The drawback of this type of method is that the researcher must select other hyperparameters, and the resulting clusters can be highly sensitive to these choices. Moreover, DBSCAN does not classify all points in the sample. Recently, Trebbi and Weese (2019) develop an interesting method to identify the number of organized insurgent groups using correlations in the timing of attacks over space. It is not obvious how to apply their method in our setting however, as identification of the number of clusters would require covariates that vary independently across our candidate characteristics, which we do not have.

of Squares (WCSS),[24] which measures how tightly packed the clusters are (i.e., how similar points within a cluster are). The second measure is the Silhouette Coefficient (SC), which measures how far apart the clusters are from one another, each for a given value of K. Ideally, the clusters should result in a low WCSS and a large SC.

Let the characteristics of candidate i be given by $\mathbf{x}_i$, and for a given value of K, denote the centroid of cluster C_k by b_k, $k = 1, ..., K$. The WCSS is then:

$$WCSS(K) = \sum_{k=1}^{K} \sum_{i \in C_k} ||\mathbf{x}_i - b_k||^2$$

This measure is shown on the first panel of Figure 1 for values of K ranging from $K = 1, ..., 20$. Notice that the $WCSS(K)$ is decreasing in K by construction - the more clusters, the more similar the members of the cluster will be. As can be seen in the figure, up to $K = 4$ each additional cluster creates large drops in $WCSS(K)$. By comparison, the gains from additional clusters are much smaller. Panels 2-4 of Figure 1 display other useful functions of $WCSS(K)$ for validating K (Makles (2012)). In panel 2 we display $\ln(WCSS(K))$, which exhibits a very similar pattern to $WCSS(K)$. In the third panel shows the η^2 coefficient, computed as $1 - \frac{WSSC(K)}{WSSC(1)}$, and we again see a similar pattern. Finally in panel 4 we have the proportional reduction of error coefficient (PRE), which is computed as $\frac{WSSC(K-1) - WSSC(K)}{WSSC(K-1)}$ and shows the additional reduction $WCSS(K)$ achieved when adding an additional cluster, from $K - 1$ to K. Here, using 4 clusters instead of 3 reduces the WSSC by 36%, while using 5 instead of 4 only causes an additional reduction of 8%. Based on these measures, $K = 4$ seems to be the appropriate number of clusters.[25]

As a second check we also consider the Silhouette Coefficient. Let $C(i)$ denote the cluster of candidate i, and define:

$$s_i = \frac{b_i - a_i}{\max\{a_i, b_i\}}$$

where

$$a_i = \frac{1}{|C(i)| - 1} \sum_{j \neq i, j \in C(i)} ||\mathbf{x}_i - \mathbf{x}_j||$$

is the average distance between i and all other points in the same cluster, and

$$b_i = \min_{k: C_k \neq C(i)} \frac{1}{|C_k|} \sum_{j \in C_k} ||\mathbf{x}_i - \mathbf{x}_j||$$

[24]The Within Cluster Sum of Squares is also often referred to as the "Inertia" score.

[25]This approach for validating the choice of K is often referred to as the "elbow method" (Thorndike (1953)).

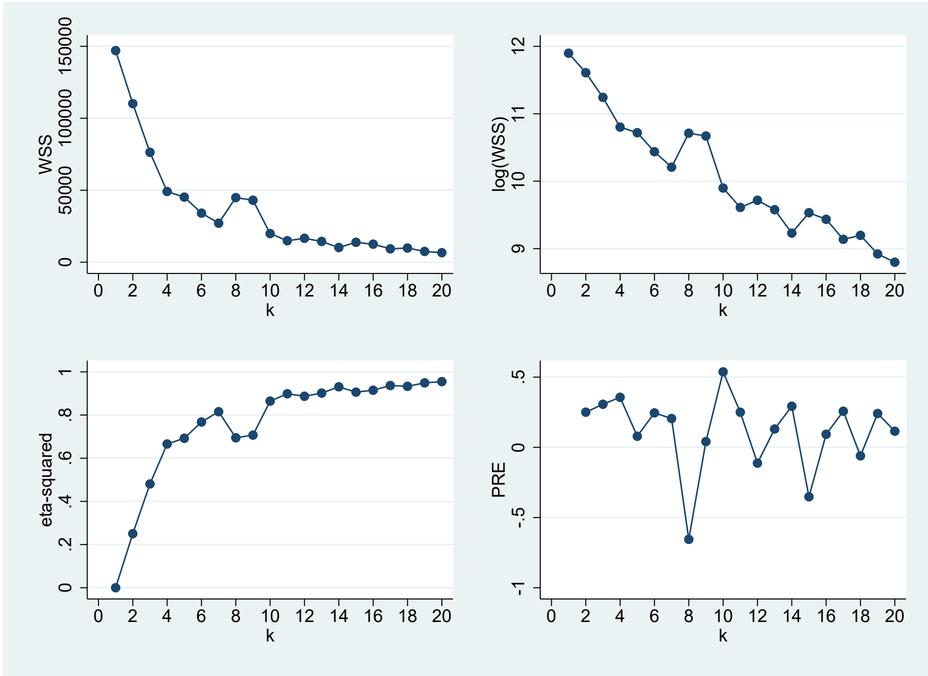

Figure 1: Diagnostics of the k-means algorithm for different values of K

is the average distance between i and the points in the cluster nearest to i other than the one it was assigned to. Notice that $s_i \in [-1, +1]$, and a value of $s_i \simeq 1$ implies that i is close to other points in its assigned cluster, and far from other clusters, while a value of $s_i \simeq -1$ implies that i is close to points in other clusters relative to points in its own. The Silhouette Coefficient is the average over all candidates in the sample:

$$SC(K) = \frac{1}{N} \sum_i S_i$$

In Figure 2 we display the Silhouette Coefficient over the same range of values of K. The largest gain in SC, by far, occurs when moving from $K = 3$ to $K = 4$, consistent with the results for $WCSS$ above. When moving from $K = 4$ to $K = 5$, SC actually decreases slightly,[26] and increases in SC are relatively small for $K > 4$.

As discussed in the next section, $K = 4$ yields candidate types that are easy to interpret. This fact, combined with the above validation checks leaves us confident in this choice and the consequent choice set for political parties.[27]

[26] Unlike $WCSS$, SC need not be monotone in K.

[27] We also experimented with $K = 5$ and $K = 6$ but found that in these cases the resulting clusters were very sensitive to the starting values of the centroids used in the algorithm. This was not the case for 4 clusters, which is stable and invariant with respect to the initialization.

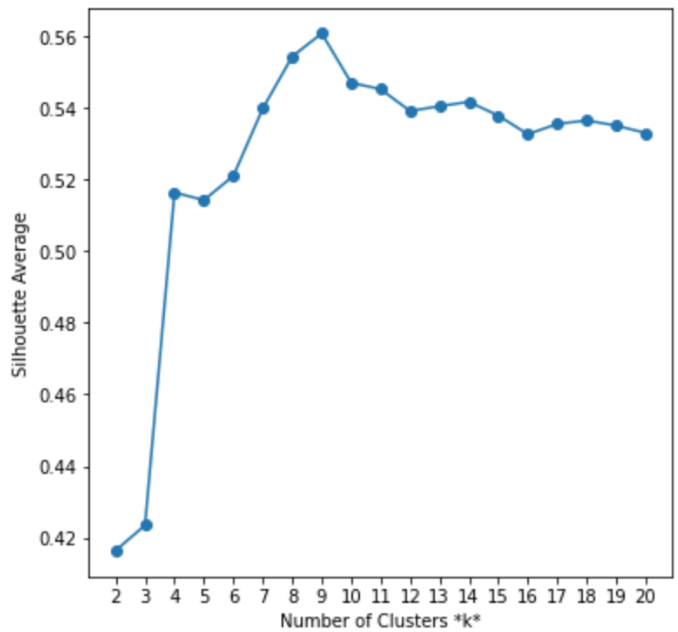

Figure 2: Silhouette Score for different values of K

5.3.3 Candidate types

Table 6 shows the centroids of each of the 4 candidate types resulting from the k-means clustering algorithm. For ease of interpretation, the table shows values on the original scale of each variable (rather than the standardized scale on which the clustering algorithm is run). The types resulting from the algorithm turn out to have fairly clear interpretations. Type 1 is an educated candidate, who has no criminal history and is non-Muslim. Type 2 is a similar candidate, but has low education. Type 3 contains all the Muslim candidates. Type 4 contains all the non-Muslim candidates with criminal history. These candidates are also richer than the other three types. For simplicity, we will refer to the four types as *educated*, *uneducated*, *Muslim*, and *criminal* types, respectively.

Table 7 shows the distribution of candidate types among all candidates, as well as the candidates of the INC and the BJP. Compared to the INC, the BJP has a somewhat higher share of uneducated types and (not surprisingly) a smaller share of Muslim types.

5.3.4 Types in the national elections

Our goal is to assign national candidates to the candidate types created from the state elections data. To do this, we first impute missing characteristics as above. There are 6581 observations in the dataset. Assets and criminal history is imputed for 3509 observations, and education is imputed for 3677. Four candidates' characteristics cannot be imputed because

Table 6: Centroids of candidate types

	Assets	Crimes	Education	Muslim
Type 1	14.35	0.00	1.00	0.00
Type 2	14.27	0.00	0.00	0.00
Type 3	14.30	0.17	0.44	1.00
Type 4	15.06	1.00	0.68	0.00

Notes: Centroids resulting from the k-means clustering algorithm. The algorithm is run on standardized variables; the table shows the centroids transformed back to the original scale for ease of interpretation.

Table 7: Distribution of candidate types in state elections

	All candidates		INC		BJP	
	N	%	N	%	N	%
Type 1	12771	34.74	1,211	49.79	1321	47.97
Type 2	14571	39.63	411	16.9	582	21.13
Type 3	4117	11.20	244	10.03	142	5.16
Type 4	5305	14.43	566	23.27	709	25.74
Total	36764	100	2,432	100	2754	100

Notes: Candidates in the state election data assigned to each type by the clustering algorithm.

there are no similar candidates in terms of gender, caste, and age bin, so we drop these from the data, resulting in 6577 observations. Candidates are assigned a Muslim indicator based on their name using the same algorithm as above.

We assign each candidate to the type that has the closest centroid. The resulting distribution of candidate types is shown in Table 8, and Tables A.1 and A.2 in the Appendix. Relative to state candidates, there are relatively more candidates of the educated type (and fewer of the uneducated type). There are also more of the criminal type, but this difference between state and national elections is less pronounced for candidates of the INC and the BJP. The share of the Muslim type is similar between state and national elections. As in the state elections, the most pronounced difference between the BJP and the INC is the former's lower share of the Muslim type and, to a lesser extent, its higher share of candidates of the uneducated type.

Table 8: Distribution of candidate types in national elections

	All candidates		INC		BJP	
	N	%	N	%	N	%
Type 1	2970	45.16	206	55.53	202	58.05
Type 2	1383	21.03	23	6.2	32	9.2
Type 3	821	12.48	36	9.7	18	5.17
Type 4	1403	21.33	106	28.57	96	27.59
Total	6577	100	371	100	348	100

Notes: Candidates in the national election data assigned to each type. Assignment is based on which type has the closest centroid.

5.4 Specification, identification, and estimation of party objectives

We focus on candidate selection by the two major party alliances in Indian politics, the UPA and the NDA. Each party's choice set contains the $K = 4$ types obtained from the clustering algorithm. Based on (2), we specify party p's objective function in constituency c as

$$U_p(a_{pc}, P(\mathbf{a}_{-p,c})) = \sum_{k=1}^{4} b_k \times \left(\sum_{\mathbf{a}_{-p,c}} w_{pc}(a_{pc}, \mathbf{a}_{-p,c}) P(\mathbf{a}_{-p,c}) \right) \mathbf{1}\{a_{pc} = k\}$$
$$+ \sum_{k=1}^{3} (c^0_{kp} + \mathbf{c}_k \mathbf{L}_{pc}) \mathbf{1}\{a_{pc} = k\} + \varepsilon_{pc}(a_{pc}), \tag{9}$$

where $\mathbf{L}_{pc}$ contains the proxies for the pool of candidate characteristics (the average of candidate education, assets, Muslim and criminal history in the assembly constituencies corresponding to constituency c). Our goal is to estimate the parameters $\theta = (\{b_k\}_{k=1}^4, \{c^0_{kp}, \mathbf{c}_k\}_{k=1}^3)$, which include parties' benefits b of increasing their probability of winning, and their costs of running different candidate types. We will refer to the costs that depend on the candidate pool, $\mathbf{c}_k$, as "recruitment costs" and the costs c^0_{kp} as "direct costs." (As will become clear below, the cost parameters are only identified for three of the four types; we use $k = 4$ as the excluded category in (9).)

As discussed above, win probabilities can be derived directly from vote shares:

$$w_{pc}(a_{pc}, \mathbf{a}_{-p,c}) = \mathbf{1}\Big\{ s_{pc}(a_{pc}, \mathbf{a}_{-p,c}) > s_{-pc}(a_{pc}, \mathbf{a}_{-p,c}) \Big\} \tag{10}$$

To compute the expected win probability $E[w_{pc}(a_{p,c}, \mathbf{a}_{-p,c}) | a_{p,c}] = \sum_{\mathbf{a}_{-p,c}} w_{pc}(a_{p,c}, \mathbf{a}_{-p,c}) P(\mathbf{a}_{-p,c})$, we need to compute the vote shares for all action profiles $(a_{pc}, \mathbf{a}_{-p,c},)$ by the players in a given constituency (16 profiles). This is done using our estimates of voter preferences. We hold fixed all exogenous variables (including non-strategic parties' choices) as well as the popularity shocks ξ_{pc}, and compute the vote shares associated with each action profile.

We provide a formal discussion of the identification of the party objective function parameters in the Appendix, but it is worth briefly discussing the sources of identification here. As is typically the case in models of discrete choice, the cost parameters are identified only up to differences with respect to a reference alternative. That is, we identify $c_k - c_4$ for $k = 1, ..., 3$ where type 4 is the reference type. The identification of these differences, for a given vector $\mathbf{b} = (b_1, ..., b_4)$ is standard (see the Appendix) and for a given value of $\mathbf{b}$ depend on the magnitude of the observed probability of selecting type k, $P(a_{p,c} = k)$ relative to the probability of selecting the reference type.

By contrast, the benefit parameters are pinned down in levels, not differences. To see this, note that given the Type 1 Extreme Value assumption we can express the relationship between choice probabilities and party payoffs as:

$$\begin{aligned} \ln(P(a_{p,c} = k)) - \ln(P(a_{p,c} = K)) &= b_k E[w_{pc}(k, \mathbf{a}_{-p,c})] - b_K E[w_{pc}(K, \mathbf{a}_{-p,c})] \\ &+ c_k - c_K + \eta_{p,c}(k) \end{aligned} \tag{11}$$

where we have assumed a single baseline type specific cost c_k for simplicity. As we can treat choice probabilities and win probabilities as known, this can be viewed as a regression of $\ln(P(a_{p,c} = k)) - \ln(P(a_{p,c} = K))$ on $E[w_{pc}(a_{p,c}, \mathbf{a}_{-p,c})|a_{p,c} = k]$ and $E[w_{pc}(a_{p,c}, \mathbf{a}_{-p,c})|a_{p,c} = K]$ where the intercept is the cost difference $c_k - c_K$. Then b_k is, loosely, identified as the covariance across constituencies between the probability of selecting type k relative to the reference option $\ln(P(a_{p,c} = k)) - \ln(P(a_{p,c} = K))$ and the expected win probability $E[w_{pc}(a_{p,c}, \mathbf{a}_{-p,c})|a_{p,c} = k]$. If the party tends to select candidate k in the constituencies where they are likely to win, b_k will be positive. Importantly, the reference parameter b_K is also identified and the logic is identical to that of b_k. See the Appendix for a more detailed discussion.

Estimation proceeds by recursively updating conditional choice probabilities using (pseudo) maximum likelihood estimates of the parameter vector θ up to convergence as in Aguirregabiria and Mira (2007). Specifically, consider an initial choice probability estimate $\hat{P}^0(\mathbf{a}_{-p,c})$. In constituency c, party p chooses a_{pc} to maximize (9). Again defining utility net of the unobservable as $\tilde{U}_p(a, P(\mathbf{a}_{-p,c}); \theta)$, the implied probability that $a_{pc} = a$ is:

$$P(a|\hat{\mathbf{P}}^0, \theta) = \frac{\exp\left\{\tilde{U}_p(a, \hat{\mathbf{P}}^0(\mathbf{a}_{-p,c}), \theta)\right\}}{\sum_{a'} \exp\left\{\tilde{U}_p(a', \hat{\mathbf{P}}^0(\mathbf{a}_{-p,c}), \theta)\right\}} \tag{12}$$

where we emphasize the fact that the choice probabilities are a function of the estimates $\hat{\mathbf{P}}^0$ as well as the parameters θ.

Denoting parties' choices observed in the data with a^*_{pc}, the log likelihood is

$$\ell(\theta,\hat{\mathbf{P}}^0) = \sum_{pc}\sum_{a} \mathbf{1}\{a = a^*_{pc}\} \ln P(a|\hat{\mathbf{P}}^0, \theta).$$

The estimates $\hat{\theta}^0$ solve

$$\max_{\theta} \ell(\theta,\hat{\mathbf{P}}^0)$$

With these estimates, we can construct new estimates of the choice probabilities as

$$\hat{P}^1(a|\hat{\mathbf{P}}^0) = \frac{\exp\left\{U_p(a, \hat{P}^0(\mathbf{a}_{-p,c}), \hat{\theta}^0)\right\}}{\sum_{a'} \exp\left\{U_p(a', \hat{P}^0(\mathbf{a}_{-p,c}), \hat{\theta}^0)\right\}}.$$

Given these new choice probabilities, the equilibrium probability that $a_{pc} = a$ in equation (12) becomes $P(a|\hat{\mathbf{P}}^1, \theta)$. In turn, this yields an updated log likelihood $\ell(\theta,\hat{\mathbf{P}}^1)$, which is maximized to obtain a new estimate $\hat{\theta}^1$. We iterate in this way until convergence. The resulting estimator, $\hat{\theta}_{NPL}$ is equivalent to the Maximum Likelihood estimator. See Aguirregabiria and Mira (2007) for details.

6 Estimation results

We first estimate a version of the model that sets parties' costs from selecting different candidates (c^0_{kp} and $\mathbf{c}_k$) to 0. This corresponds to the "standard" approach where parties only care about the probability of winning, though we allow for the fact that the valuation of this probability is affected by the candidate's type.

The parameter estimates are in column (1) of Table 9. These suggest significant differences between parties' value of winning with different candidate types. For example, increasing the probability of winning is 7 times more valuable when the party is represented by an educated type (Type 1) than when it is represented by a criminal type (Type 4). However, this model where parties care *only* about their probability of winning does not fit the data well relative to a model that allows for party preferences as we see in columns 2 and 3 where we introduce the direct costs c^0_{kp} and recruitment costs $\mathbf{c}_k$ of choosing specific types (by including type and party dummies as well as the state-candidate characteristics $\mathbf{L}_{pc}$). Note that only the relative magnitude of these cost parameters are identified. We set the criminal type (Type 4) as the excluded category, so that all cost estimates represent parties' direct payoffs relative to this type. The estimates show that costs are important in explaining parties' choices of which candidates to run. This is reflected in the chi-square tests of joint significance of the cost parameters at the bottom of columns 2 and 3. The test statistics are large (p-value smaller than 0.000) in all cases. It is worth noting that the benefit parameters remain jointly significant ($p < 0.000$) once all cost parameters are included in Column 3 as well.

Including these payoff components substantially improves the model's fit, as the Log Likelihood increases by 24% as we move from Column 1 to Column 3. See Section 5.1 below for a more detailed evaluation of model fit.[28] In these elections, parties' objective function when picking candidates is not restricted to their probability of winning.

The recruitment cost parameters $\mathbf{c}_k$ mirror the patterns seen in Table 2: a higher prevalence of some candidate characteristic in a party's candidate pool (as measured using the state election data) lowers the party's cost of selecting a candidate type with that characteristic. For example, the estimates show that a higher prevalence of educated candidates in the relevant pool increases a party's payoff from choosing the educated type (Type 1) relative to the other types ($c_1^{educ} > 0 > c_2^{educ}, c_3^{educ}$). Similarly, the negative estimates of c_k^{crime} for $k = 1, 2, 3$ indicate that a higher prevalence of criminal candidates in the pool increases the party's payoff from choosing a criminal type (Type 4). The supply of candidate characteristics available to parties appears to affect who they choose to run.

Column 3 shows that parties' costs of running specific candidates are not restricted to the recruitment costs. According to these estimates, the NDA has a lower direct cost from running a criminal type (Type 4) relative to a low-educated or a Muslim type (Types 2 and 3). The UPA has a lower cost of running a Muslim type than running a criminal type. Both parties incur similar direct costs from running an educated type (Type 1) or running a criminal type.

The idea that parties (particularly the NDA, according to our estimates) can have relatively low costs of running criminal candidates is in line with Vaishnav (2017)'s argument that these candidates are willing to fund their own campaigns (perhaps circumventing campaign finance laws), which makes them cheaper for the parties. Other sources of these direct payoffs could include benefits to party leaders from co-opting powerful local crime bosses.

Perhaps most surprisingly, we estimate that $b_4 < 0$, indicating that parties do *not* like winning with criminal candidates. This is in stark contrast to the other three candidate types, all of which yield positive payoffs to parties as their probability of winning increases.

[28] In the Appendix we provide more results illustrating how the fit improves as we move from a model where parties care only about voter preferences (Column 1) to one where they are allowed to have their own preferences over candidates.

Intuitively, the finding that $b_4 < 0$ suggests that parties do not choose criminal candidates often enough given voters' preference for this type of candidate, particularly in constituencies where, based on our estimates, choosing a criminal candidate would yield a large probability of winning. Instead, parties choosing criminal candidates is rationalized by $c_4 > 0$, a direct benefit to the party from selecting the criminal type.

Why do parties not select criminal candidates when voters have a strong preference for them? A possible answer is that parties care about future policies and their reputation, and elected criminals could be damaging to these. Indeed, Vaishnav (2016) argues that voters tend to prefer criminal candidates in places where the rule of law is weak (or unevenly enforced) and social divisions are highly salient, because here criminals can guarantee security and provide services to large groups of voters. In doing so, however, criminal politicians tend to exploit local divisions and have perverse incentives once in office, which can be damaging to their party.

A voter preference for criminals together with parties' aversion to winning with them creates the possibility for equilibrium inefficiency (from the parties' perspective). To see this, suppose the NDA has a high probability of running a criminal type. The UPA can also run a criminal type, which delivers negative payoffs conditional on winning (but yields positive direct payoffs). Or, it can run a type that would deliver positive payoffs conditional on winning - but this type has a low probability of winning against the NDA's criminal type. If the difference in direct payoffs is large enough, the UPA will be likely to run a criminal as well. Thus, in equilibrium the parties may run criminals more often than they would like to.

This possibility is supported by the fact that, at baseline, the correlation between the two parties' choice probability of a criminal type is positive (0.29), while the correlation between one party's CP for a criminal type and the other party's CP for any other type is always negative (ranging between -0.07 and -0.29). See Table A.3 in the Appendix.

To interpret the magnitude of the estimates, a useful benchmark is parties' payoff from winning. For example, the NDA's cost of running a Muslim candidate relative to a criminal (-1.84) is larger in absolute value than the benefit of winning with such a candidate (1.42). This is not surprising given the Hindu nationalist profile of the BJP, the NDA's leading party. For the UPA, the cost of running a Muslim candidate is much smaller in absolute value (-1.14).

Figure 3 plots the recruitment costs for each type, and compares them to the direct costs (indicated with vertical lines). Interestingly, recruitment costs tend to be negative, indicating that in most constituencies, parties incur relatively low costs from selecting criminal types compared to other types due to the ample supply of criminality in the candidate pool. Overall, recruiting costs tend to be lower in magnitude than the corresponding direct cost.

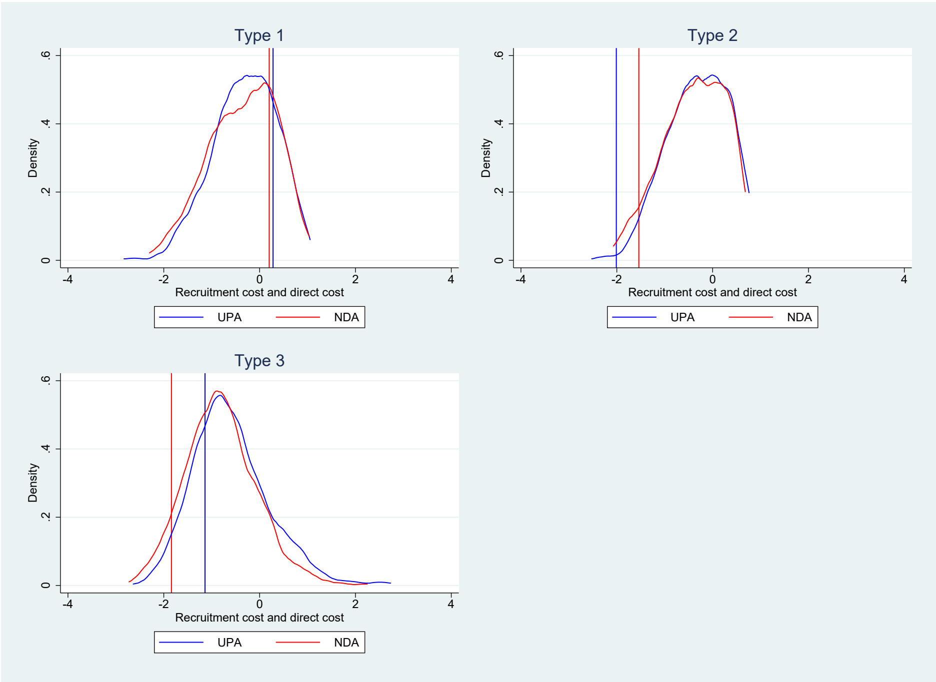

Figure 3: Distribution of each party's costs of different types

Based on column 3 of Table 9. Kernel density plots of recruitment costs; direct costs indicated with vertical lines.

Table 9: Party objective function estimates

	(1)	(2)	(3)
b_1	2.10	1.11	0.72
	(0.26)	(0.27)	(0.27)
b_2	-0.12	0.31	0.97
	(0.32)	(0.34)	(0.37)
b_3	0.57	0.81	1.42
	(0.41)	(0.43)	(0.46)
b_4	0.29	-0.69	-0.93
	(0.13)	(0.17)	(0.19)
$c^0_{1,NDA}$			0.20
			(0.21)
$c^0_{2,NDA}$			-1.54
			(0.32)
$c^0_{3,NDA}$			-1.84
			(0.34)
$c^0_{1,UPA}$			0.28
			(0.21)
$c^0_{2,UPA}$			-2.01
			(0.35)
$c^0_{3,UPA}$			-1.14
			(0.33)
c^{educ}_1		0.79	0.49
		(0.20)	(0.21)
c^{educ}_2		-1.08	-0.04
		(0.28)	(0.35)
c^{educ}_3		-1.10	-0.11
		(0.29)	(0.36)
c^{crime}_1		-1.05	-1.08
		(0.14)	(0.14)
c^{crime}_2		-1.39	-1.06
		(0.22)	(0.25)
c^{crime}_3		-1.15	-0.70
		(0.21)	(0.24)
c^{asset}_1		-0.07	-0.19
		(0.14)	(0.16)
c^{asset}_2		-0.77	0.18
		(0.20)	(0.27)
c^{asset}_3		-1.05	-0.44
		(0.21)	(0.27)
c^{Muslim}_1		-0.11	-0.04
		(0.24)	(0.23)
c^{Muslim}_2		0.28	-0.05
		(0.34)	(0.42)
c^{Muslim}_3		1.60	1.49
		(0.29)	(0.31)
Log likelihood	-1144.4	-926.0	-869.6
Joint significance of b's	109.54	57.78	55.79
Joint significance of c's	-	304.64	376.60

Notes: Number of markets: 434.

7 Model validation and fit

To explore the model's ability to fit the data, we first use the estimated model to simulate party choices in every constituency. We repeat this 100 times and compare the average over simulations with the actual choices observed in the data for each type and each party. The result is in Table 10. The model performs remarkably well; at the level of the alliance the model prediction is always within one (on average) of the actual number of candidates selected of any given type.

Table 10: Model fit

Type	UPA actual	UPA predicted	NDA actual	NDA predicted	All actual	All predicted
1	217	217.85	229	229.7	446	447.12
2	24	23.65	43	43.66	67	67.31
3	49	49.12	22	21.43	71	70.55
4	144	143.38	140	139.64	284	283.02

Notes: Number of candidates of each type observed in the data and predicted by the model.

On the one hand it is encouraging to have a model with this ability to fit the data. On the other hand we should be concerned about the possibility that we have over-fit the data. Specifically, we may have parameterized our model to the point that we can predict outcomes in this particular sample with high accuracy (at least on average) but if we were faced with a new data set drawn from the population it would not fit well.

To address this possibility and further study the model's validity, we use a simple cross-validation procedure. We hold out 20% of the sample, estimate the model using the remaining 80%, and evaluate the model's fit on the held out sample. To do this, we use the estimates obtained from the 80% estimation sample to solve for the equilibrium in each constituency in the hold-out sample, and use these to make our predictions. If the model is able to predict outcomes in the held-out sample well, we should be confident that we have a model with high predictive accuracy (and thus suitable for considering counterfactual experiments) and is externally valid (not prone to over-fitting). To avoid the possibility of picking a fortuitous split of the data that yields a good fit, we repeat this 5 times and take an average (essentially a k-fold cross validation, with $k = 5$).

The results are in Table 11. While there is variation across the folds in predictive ability, on average over the k-folds the model does just as well in predicting outcomes as in Table 10.

Table 11: Model fit using k-fold cross validation

Type	UPA actual	UPA predicted	NDA actual	NDA predicted	All actual	All predicted
1	41.4	41.34	43.4	43.822	84.8	85.162
2	4.8	4.212	8.4	8.328	13.2	12.54
3	9.2	9.446	4.4	4.252	13.6	13.698
4	27.6	28.002	26.8	26.598	54.4	54.600

Notes: Number of candidates of each type observed in the data and predicted by the model, using k-fold cross validation as described in the text.

8 Policy experiment: banning criminal candidates

What is the impact of banning candidates with a criminal history from contesting the election? As discussed in the Introduction, this is a relevant policy question in many settings. For example, in Lily Thomas v. Union of India (2013) the Supreme Court ruled that individuals who served two or more years in prison were barred from running in state and national elections for a period of six years.

To model this, we consider a counterfactual scenario where we make it prohibitively costly to choose the criminal type (Type 4) for both parties.[29] Given our parameter estimates, we compute a new equilibrium, and study the resulting choice probabilities of the parties and associated winning probabilities of the remaining three candidate types.[30]

8.1 Changes in candidate characteristics

A first observation is that, because candidates are bundles of correlated characteristics, eliminating a candidate type directly affects the distribution of characteristics among candidates. The criminal type is relatively wealthier, and comes from the religious majority (non-Muslim). Eliminating these candidates may therefore raise the share of less wealthy and minority candidates. Such changes may be unintended side effects of a policy of banning candidates with a criminal history.

Understanding the impact of banning criminal candidates *in equilibrium* is complicated by the fact that parties respond by adjusting their candidate choices. To get a sense of what this entails, consider first banning the criminal Type 4 for one party only. For concreteness, we simulate a ban for the UPA (the symmetric exercise of banning the criminal type only

[29] Although Type 3 also contains some candidates with criminal history, Type 4 candidates *always* have a criminal history (Table 6). They are also wealthier, and thus more closely match the kind of criminal candidates who are often considered problematic in the Indian context (see Vaishnav (2017)).

[30] In the main analysis, we consider removing the criminal type only for the UPA's and the NDA's choice set in the national election. We discuss removing these candidates from the state election pool or from third parties' choice sets below.

for the NDA yields very similar patterns). Table 12 shows the NDA's choice probabilities at baseline, as well as the counterfactual change in its choice probabilities following the ban. This yields two noteworthy patterns.

First, the NDA's choice probability of the criminal type weakly decreases in *every* constituency. The average decrease in this CP is 1 percentage point (3 percentage points if zeros are ignored), relative to a mean of 32%. Once its opponent is prevented from running a criminal, the NDA's incentive to run a criminal declines. This complementarity is due to the fact that, as seen above, voters like criminals, but parties do not like winning with them. At baseline, once the UPA chooses a criminal, a type that would deliver higher payoffs for the NDA conditional on winning would be unlikely to win. This increases the NDA's incentive to choose criminals. When the UPA is prevented from choosing a criminal, this incentive is attenuated, leading the NDA to choose criminal candidates less often.

Second, as the NDA substitutes to non-criminal types (Type 1-3), its CPs for each of the other types can increase *or decrease*. This is contrary to the simple intuition from standard Logit models where reducing the choice set could never reduce the probability of a remaining option. The difference is due to the fact that here parties play best responses to each-others' strategies. To illustrate with an example that ignores incomplete information, take a constituency where the NDA ran a Type 1 candidate and the UPA a Type 4 candidate. Losing the option of running a Type 4 candidate, the UPA might switch to Type 2. If the NDA's best response is to switch to Type 3, then all else equal the share of Type 2 and 3 candidates would increase, but the share of Type 1 would decrease. This illustrates that, in general, a policy of banning criminal types for both parties can result in an increase in some candidate types and a decrease in others.

Table 12: Choice probabilities for the NDA with criminal ban for the UPA only

Type	Baseline	Change			
	Mean	Min	Max	Mean	Mean if ΔCP4<0
1	0.53	-0.07	0.09	0.01	0.02
2	0.10	-0.03	0.07	0.00	0.01
3	0.05	-0.01	0.08	0.00	0.01
4	0.32	-0.12	0.00	-0.01	-0.03

Notes: The table shows the NDA's CPs for the different types at baseline, and how they change when the criminal type is banned for the UPA. For changes in CPs, the values shown are the min, max and mean across constituencies, as well as the mean across the 168 constituencies where the change in the criminal type's CP was nonzero.

Turning to the policy experiment of a ban of the criminal type for both parties, the first two columns of Table 13 compare the average choice probabilities (i.e., the predicted share of each candidate type among the candidates contesting the election) in the no-criminal counterfactual equilibrium and the baseline equilibrium. The full distribution of these probabilities is shown on Figure 4.

Table 13: Choice probabilities and winning probabilities with and without Type 4

	Avg. choice probability		Avg. win probability	
	baseline	counterfactual	baseline	counterfactual
All candidates				
Type 1	51.4	76.0	17.7	20.9
Type 2	7.7	11.7	20.6	24.2
Type 3	8.2	12.3	9.5	11.6
Type 4	32.7	0.0	57.2	
UPA				
Type 1	50.0	74.9	14.0	17.0
Type 2	5.5	8.4	16.8	20.2
Type 3	11.3	16.7	7.8	9.6
Type 4	33.2	0.0	48.6	
NDA				
Type 1	52.8	77.0	21.5	24.8
Type 2	9.9	15.1	24.4	28.1
Type 3	5.1	7.9	11.1	13.7
Type 4	32.3	0.0	65.8	

Notes: Type 1: "educated," Type 2: "uneducated," Type 3: "Muslim," Type 4: "criminal". Values shown are the averages across all the constituencies in the data. Average winning probability is the probability that a type would win conditional on being chosen.

We find that the choice probabilities of candidate types 1-3 all increase. As hihlighted above, this is *not* simply mechanical, because unlike in a simple Logit model, parties play best responses to each-other's strategies. For each type, we find that decreases in choice probabilities occur in less than 2% of cases.

In absolute terms, the increase in average choice probabilities is largest for the educated type (Type 1), whose expected share increases by 24.6 percentage points, from 51.4 to 76 percent (Table 13). However, changes for the other two types are similar in relative terms. Eliminating criminal candidates increases the share of uneducated candidates (Type 2) by 52.2 percent (from 7.7 to 11.7). It also increases the share of minority (Muslim) candidates (Type 3) by 50 percent (from 8.2 to 12.3). The latter effect is larger for the NDA, where the share of Muslim candidates was relatively low.

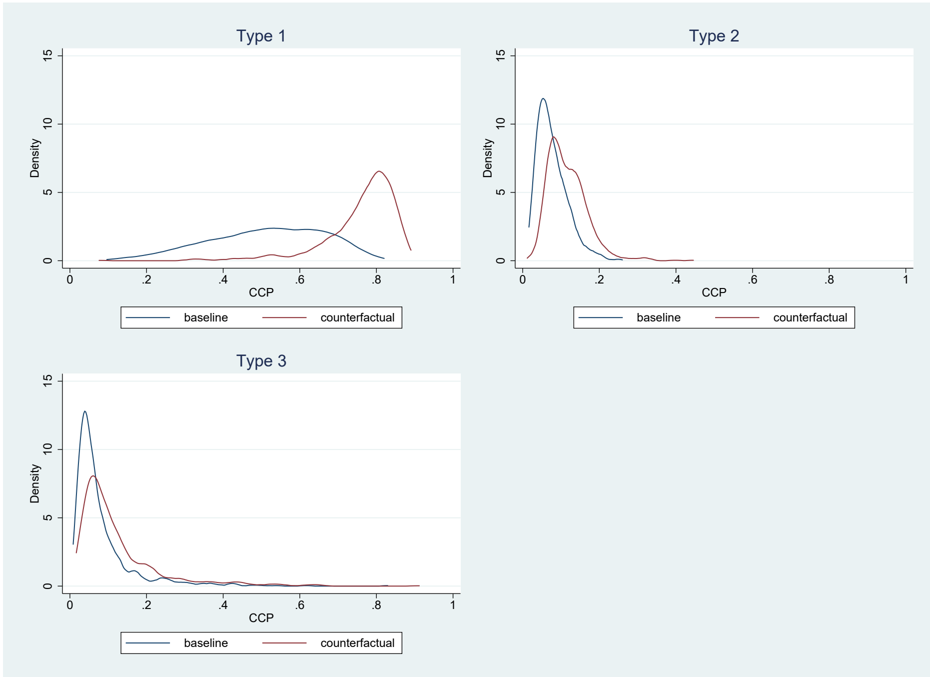

Figure 4: Distribution of choice probabilities by type
Kernel density plots of choice probabilities in the baseline and the counterfactual.

8.2 Changes in different types' winning probabilities

As shown in Table 13, these changes also impact the viability of political viability of different candidate types, captured by the expected winning probabilities $E_P[w_p(a_p, \mathbf{a}_{-p})|a_p]$ (the probability that a type would win conditional on being selected). For example, the winning probability of the uneducated type rises from 20.6% in the baseline to 24.2% in the counterfactual. Figure 5 shows the distribution of these winning probabilities by type, in the baseline and the counterfactual (conditional on $E_P[w_p(a_p, \mathbf{a}_{-p})|a_p]$). As can be seen, in each case the increase in winning probabilities is due to a shift from the middle of the distribution to higher values (i.e., it is not due to changes in small probabilities).

8.3 Changes in parties' winning probabilities and payoffs

Figure 6 shows the changes in the two parties' probability of winning following a ban on criminal types. In most cases we find that both parties' probability of winning falls. In the average constituency, the probability that the UPA wins drops from 0.23 to 0.17, and the probability that the NDA wins from 0.34 to 0.25. The same is true for the probability that either party wins: removing the criminal type raises the probability that a third party wins

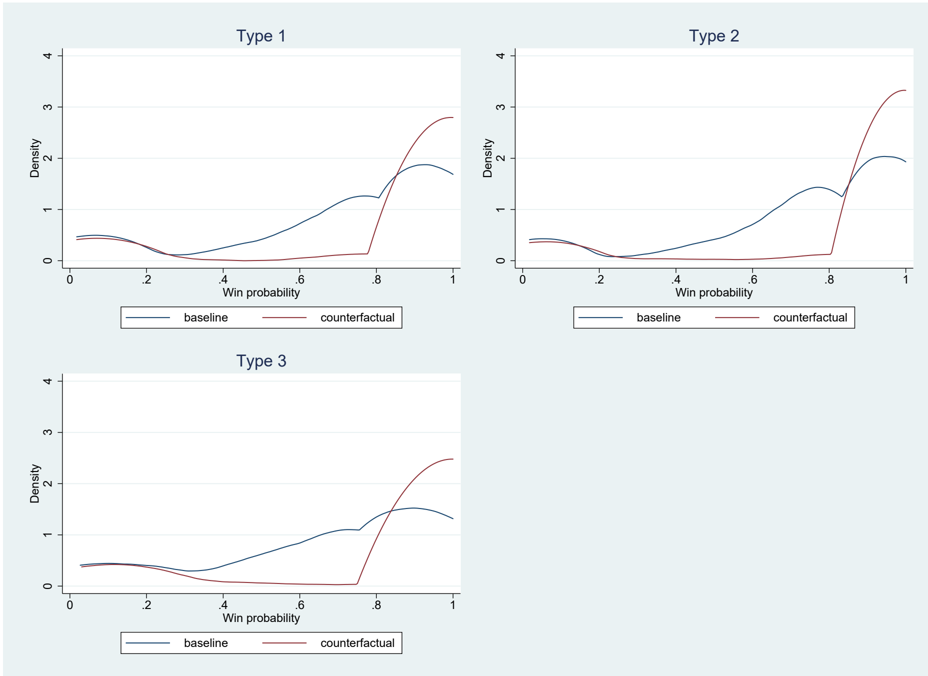

Figure 5: Distribution of each type's nonzero winning probabilities

Kernel density plots of each type's nonzero winning probabilities in the baseline and the counterfactual.

in 239 markets, and increases it in only 9.[31]

Once criminal types are not available, voters appear to find the main parties less attractive relative to third parties. To quantify this, we use our voter preference estimates to compute third parties' average vote shares for all possible choice profiles of the two main parties. In the seven profiles where at least one of the main parties chooses a criminal type, the average vote share of third parties is 10% (ranging from 7-12% across profiles). In the nine profiles where neither main party chooses a criminal type, this vote share increases to 17% (with a range of 15-20% across profiles). While in equilibrium the parties can mitigate the loss in votes by switching candidates, the results from our counterfactual exercise suggest that they cannot fully offset this decrease in their popularity, and their vote share goes down.

Computing the changes in parties' expected equilibrium payoffs yields the distributions shown on Figure 7. On average, both the UPA's and the NDA's payoffs decline when the

[31] It is important to note that the reduction in winning probabilities for the two major parties is *not* due to voters' switching to third parties who run criminals. First, we obtain similar patterns when we restrict attention to the 82 markets where none of the third parties runs a criminal: here the probablity that the UPA wins drops from 0.34 to 0.29, and the probability that the NDA wins from 0.41 to 0.35. Second, we also ran a set of counterfactuals where we removed all third parties running criminals before imposing the ban on criminal types for the two main parties. We again found similar declines in winning probabilities for both the UPA (from 0.30 to 0.26) and the NDA (from 0.43 to 0.37).

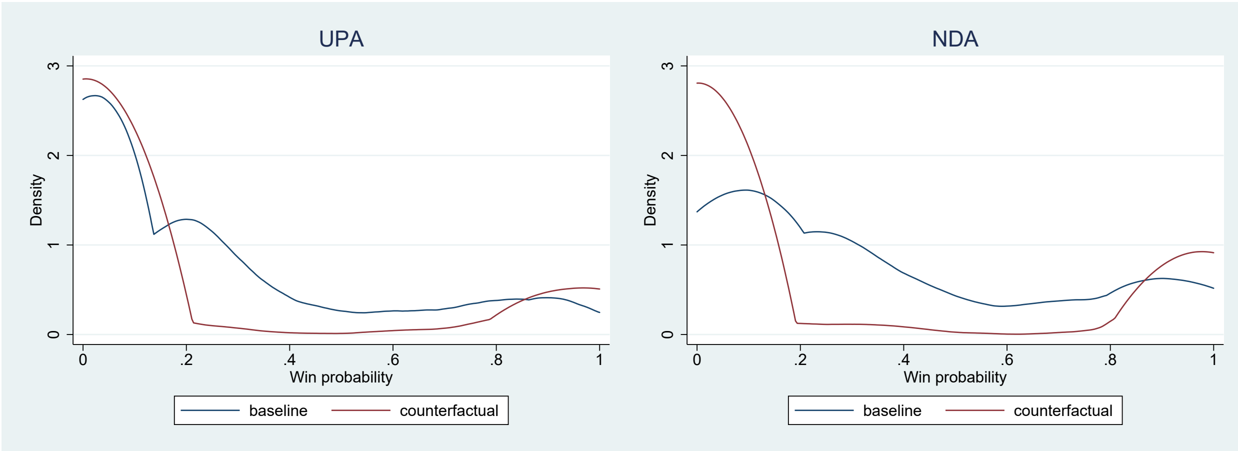

Figure 6: Distribution of parties' winning probabilities
Kernel density plots of parties' unconditional winning probabilities in the baseline and the counterfactual.

criminal type is banned. However, there are also many constituencies where payoffs rise - this is the case for in 60% of the constituencies for the NDA and in 47% for the UPA. Furthermore, changes in the two parties' payoffs are positively correlated (the correlation is 0.26). In 32% of the constituencies both parties gain, while in 25% both parties lose.

The possibility that both parties may gain from a ban on criminal types arises from the equilibrium inefficiencies described above. According to our model, in some cases a party may be forced to choose a criminal because its opponent chooses a criminal. In this case a candidate who would deliver higher payoffs conditional on winning would be unlikely to win, so that neither party has an incentive to choose a non-criminal type. Banning the criminal type removes this inefficiency and allows parties to profitably choose different candidates. This can increase parties' payoffs - even though it lowers their probability of winning.

These findings imply that, in some circumstances, the main Indian parties may be willing to support a ban on criminal candidates.

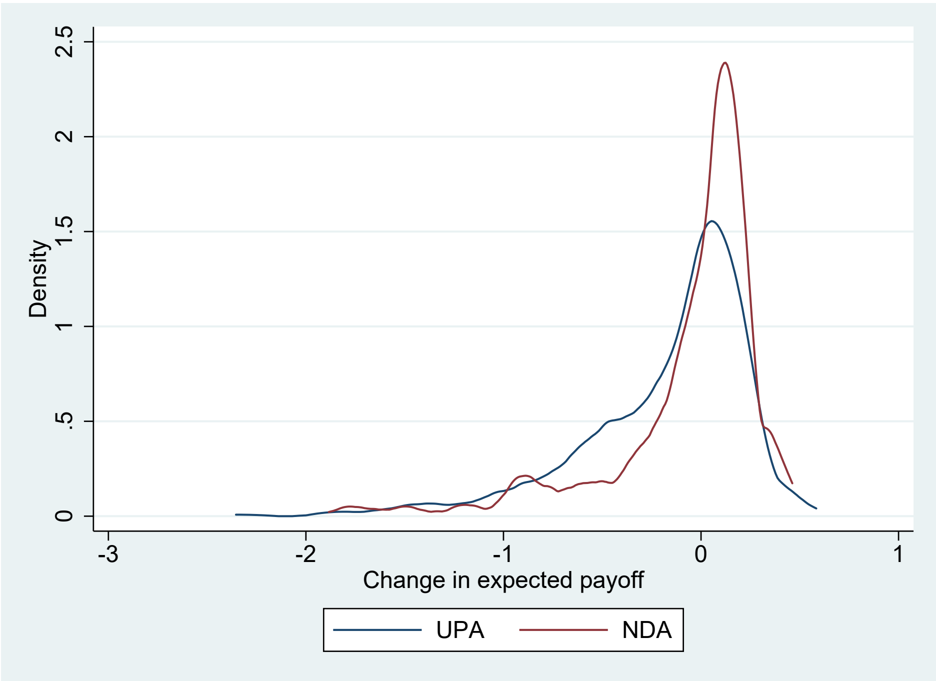

Figure 7: Distribution of the change in parties' payoffs
Kernel density plots of the change in parties' expected payoffs (counterfactual payoff minus baseline payoff).

9 Conclusion

We estimate a model of candidate selection by political parties to study why parties in a representative democracy select the candidates they do. Our setting is India, which is notorious for having high rates of criminality among elected officials. We combine a rich demand side specification of voter preferences with a supply side game between parties that incorporates direct payoffs from candidate selection. A machine learning algorithm is used to assign candidates to types base on detailed information on their characteristics.

We find that while voter preferences influence party decisions, party preferences over candidate types are the main force that shapes candidate selection. Although parties prefer to win with noncriminal candidates, they often select criminal candidates because of the direct utility they provide, perhaps in the form of networks and party finances. This can create inefficiency in equilibrium - the expectation that their rival will run a criminal candidate makes it difficult for parties to win without running a criminal candidate in response.

We use the estimated model to study the consequences of a ban on criminal candidates, an idea that India's political establishment has long debated but never implemented. We find that a ban of this nature reduces the chances that India's major national parties win the election, as the ban causes voters to consider third party alternatives. In other words, the appeal of the major parties is inextricably linked to this type of candidate. At the same time, banning criminal candidates can increase parties' payoffs by removing the equilibrium inefficiency.

Our paper may provide insights for other countries where crime in politics is a salient problem, and more generally a method to estimate the factors that guide parties' selection of candidates with particular characteristics.

References

Aguirregabiria, Victor, and Pedro Mira. 2007. "Sequential estimation of dynamic discrete games." *Econometrica*, 75(1): 1–53.

Asher, Sam, Tobias Lunt, Ryu Matsuura, and Paul Novosad. 2020. "Development Research at High Geographic Resolution: An Analysis of Night Lights, Firms, and Poverty in India using the SHRUG Open Data Platform." *working paper*.

Athey, Susan, and Guido W Imbens. 2019. "Machine learning methods that economists should know about." *Annual Review of Economics*, 11: 685–725.

Bandiera, Oriana, Andrea Prat, Stephen Hansen, and Raffaella Sadun. 2020. "CEO behavior and firm performance." *Journal of Political Economy*, 128(4): 1325–1369.

Banerjee, Mukulika. 2017. *Why India Votes?* Routledge.

Berry, Steven, James Levinsohn, and Ariel Pakes. 1995. "Automobile prices in market equilibrium." *Econometrica*, 63(4): 841–890.

Besley, Tim. 2005. "Political Selection." *Journal of Economic Perspectives*, 19(3): 43–60.

Besley, Timothy, Olle Folke, Torsten Persson, and Johanna Rickne. 2017. "Gender quotas and the crisis of the mediocre man: Theory and evidence from Sweden." *American economic review*, 107(8): 2204–2242.

Casey, Katherine, Abou Bakarr Kamara, and Niccoló F. Meriggi. 2021. "An experiment in candidate selection." *American Economic Review*, 111(5): 1575–1612.

Dal Bó, Ernesto, and Frederico Finan. 2018. "Progress and perspectives in the study of political selection." *Annual Review of Economics*, 10: 541–575.

Dal Bó, Ernesto, Frederico Finan, Olle Folke, Torsten Persson, and Johanna Rickne. 2017. "Who becomes a politician?" *The Quarterly Journal of Economics*, 132(4): 1877–1914.

Dar, Aaditya. 2019. "Parachuters vs. Climbers: Economic Consequences of Barriers to Political Entry in a Democracy."

Farooqui, Adnan, and Eswaran Sridharan. 2014. "Incumbency, internal processes and renomination in Indian parties." *Commonwealth & Comparative Politics*, 52(1): 78–108.

Galasso, Vincenzo, and Tommaso Nannicini. 2011. "Competing on good politicians." *American political science review*, 105(1): 79–99.

Gandhi, Amit, and Jean-François Houde. 2019. "Measuring substitution patterns in differentiated-products industries." *NBER Working paper w26375*.

Gulzar, Saad, Zuhad Hai, and Binod Kumar Paudel. 2021. "Information, candidate selection, and the quality of representation: Evidence from Nepal." *The Journal of Politics*, 83(4): 1511–1528.

Hamilton, Barton H., Andrés Hincapié, Robert A. Miller, and Nicholas W. Papageorge. 2021. "Innovation and Diffusion of Medical Treatment." *International Economic Review*, 62(3): 953–1009.

Heath, Oliver, and Adam Ziegfeld. 2022. "Why So Little Strategic Voting in India?" *American Political Science Review*, 116(4): 1523–1529.

Iaryczower, Matias, Galileu Kim, and Sergio Montero. 2023. "Representation Failure." *working paper.*

Kapoor, Sacha, and Arvind Magesan. 2018. "Independent candidates and political representation in India." *American Political Science Review*, 112(3): 678–697.

Makles, Anna. 2012. "Stata tip 110: How to get the optimal k-means cluster solution." *The Stata Journal*, 12(2): 347–351.

Mattozzi, Andrea, and Antonio Merlo. 2015. "Mediocracy." *Journal of Public Economics*, 130: 32–44.

Nevo, Aviv. 2000. "A practitioner's guide to estimation of random-coefficients logit models of demand." *Journal of Economics & Management Strategy*, 9(4): 513–548.

Nevo, Aviv. 2001. "Measuring market power in the ready-to-eat cereal industry." *Econometrica*, 69(2): 307–342.

Olea, José Luis Montiel, and Carolin Pflueger. 2013. "A Robust Test for Weak Instruments." *Journal of Business & Economic Statistics*, 31(3): 358–369.

Paiva, Natália, Juliana Sakai, and Lauren Schoenster. 2014. *Quatro em cada dez candidatos a governador têm ocorrências na Justiça ou nos Tribunais de Contas.* Transparencia Brasil, https://www.transparencia.org.br/publicacoes.

Prakash, Nishith, Marc Rockmore, and Yogesh Uppal. 2019. "Do criminally accused politicians affect economic outcomes? Evidence from India." *Journal of Development Economics*, 141: 102370.

Roy, Ramashray. 1966. "Selection of Congress candidates. I: The formal criteria." *Economic and Political Weekly*, 833–840.

Thorndike, Robert. 1953. "Who belongs in the family?" *Psychometrika*, 18(4): 267–276.

Trebbi, Francesco, and Eric Weese. 2019. "Insurgency and small wars: Estimation of unobserved coalition structures." *Econometrica*, 87(2): 463–496.

Ujhelyi, Gergely, Somdeep Chatterjee, and Andrea Szabó. 2021. "None of the Above: Protest Voting in the World's Largest Democracy." *Journal of the European Economic Association*, 19(3): 1936–1979.

Vaishnav, Milan. 2016. "Why Voters Sometimes Prefer Criminals as Candidates." *Governance*, 29(4): 459–461.

Vaishnav, Milan. 2017. *When crime pays: Money and muscle in Indian politics.* Yale University Press.

Appendix to "Candidate selection by parties: Crime and politics in India"

Arvind Magesan
Economics Department, University of Calgary

Andrea Szabó
Economics Department, University of Houston

Gergely Ujhelyi
Economics Department, University of Houston

June 16, 2023

Contents

1 Additional tables

Table A.1: Distribution of candidate types in national elections by year

	Type 1	Type 2	Type 3	Type 4	Total	N
2009	46.12	22.05	12.54	19.3	100	3207
2014	44.24	20.06	12.43	23.26	100	3370
Total	45.16	21.03	12.48	21.33	100	6577

Notes: Type 1: educated, Type 2: uneducated, Type 3: Muslim, Type 4: criminal

Table A.2: Distribution of candidate types in national elections by state

	Type 1	Type 2	Type 3	Type 4	Total	N
Andhra Pradesh	50.92	20.94	10.68	17.45	100	487
Assam	55.73	13.02	23.96	7.29	100	192
Bihar	38.17	18.17	13.17	30.49	100	820
Gujarat	40.67	22.33	13.33	23.67	100	300
Haryana	52.34	23.83	3.4	20.43	100	235
Jharkhand	42.29	25.69	9.88	22.13	100	253
Karnataka	48.31	19.85	14.23	17.6	100	534
Kerala	25.66	14.16	38.05	22.12	100	113
Madhya Pradesh	50.34	18.37	9.52	21.77	100	147
Maharashtra	39.6	20.47	15.64	24.3	100	889
Odisha	52.23	16.15	7.22	24.4	100	291
Rajasthan	48.22	27.41	10.15	14.21	100	394
Tamil Nadu	50.39	21.34	7.59	20.68	100	909
Uttar Pradesh	43.76	23.79	11.16	21.29	100	681
West Bengal	41.87	23.19	19.28	15.66	100	332
Total	45.16	21.03	12.48	21.33	100	6577

Notes: Type 1: "educated," Type 2: "uneducated," Type 3: "Muslim," Type 4: "criminal"

Table A.3: Correlation of estimated CPs

		UPA			
		Type 1	Type 2	Type 3	Type 4
NDA	Type 1	0.27	0.14	0.06	-0.29
	Type 2	0.13	0.25	-0.12	-0.08
	Type 3	0.06	-0.08	0.21	-0.16
	Type 4	-0.27	-0.16	-0.07	0.29

Notes: Type 1: "educated," Type 2: "uneducated," Type 3: "Muslim," Type 4: "criminal"

2 Identification of Model Parameters

Here we formally establish identification of the parameters of the model of candidate selection and discuss the intuition of the identification results. Throughout, we assume that choice probabilities and win probabilities are known to the researcher - these are estimated in a first stage. Let the probability that party $i \in \{1, 2\}$ chooses action $a_i = k$ for $k = 1, ..., K$ given observable payoff variables $\mathbf{z}$ (i.e., constituency characteristics) be given by $P_i(k, \mathbf{z})$, and write expected winning probability of party i as:

$$w_i^P(k, \mathbf{z}) = E_i[w_i(a_i, a_{-i}, \mathbf{z})|a_i = k] \tag{1}$$

where the expectation $E_i[w_i(a_i, a_{-i}, \mathbf{z})|a_i = k]$ is an integration over a_{-i} using player $-i$'s choice probability (see Section 3).

We establish identification in the baseline model with type specific benefit parameters $\mathbf{b} = (b_1, b_2, ..., b_K)'$ and type specific costs $\mathbf{c} = (c_1, c_2, ..., c_K)'$ as most of the intuition can be gleaned from this case, and allowing for additional cost parameters as in our full model does not substantially change the identification argument.

Player i's choice probability satisfies:

$$P_i(k, \mathbf{z}) = \Lambda\big(b_k \times w_i^P(k, \mathbf{z}) + c_k\big) \tag{2}$$

where, given our assumption about the error distribution:

$$\Lambda\big(b_k \times w_i^P(k, \mathbf{z}) + c_k\big) = \frac{\exp\left\{b_k \times w_i^P(k, \mathbf{z}) + c_k\right\}}{\sum_{k'} \exp\left\{b_k \times w_i^P(k', \mathbf{z}) + c_k'\right\}} \tag{3}$$

As the argument for identification is symmetric across players, we drop the i subscript in what follows for expositional purposes.

Inverting the choice probability gives:

$$\begin{aligned}\Lambda^{-1}\big(P(k,\mathbf{z})\big) &= \ln\big(P(k,\mathbf{z})\big) - \ln\big(P(K,\mathbf{z})\big) \\ &= b_k \times w^P(k,\mathbf{z}) - b_K \times w^P(K,\mathbf{z}) + c_k - c_K\end{aligned} \tag{4}$$

where we have taken type K as the reference type.

Before discussing full identification of the vectors $\mathbf{b}$ and $\mathbf{c}$, to build intuition let's first consider the case where the preference for winning is common across candidate types: $b_k = b$ for all k. Then we have:

$$\Lambda^{-1}\big(P(k,\mathbf{z})\big) = b \times \big(w^P(k,\mathbf{z}) - w^P(K,\mathbf{z})\big) + c_k - c_K \tag{5}$$

Define $\Delta_w^P(k,\mathbf{z}) \equiv w^P(k,\mathbf{z}) - w^P(K,\mathbf{z})$. The difference $\Delta_w^P(k,\mathbf{z})$ represents the *increased expected probability of winning when selecting type k relative to the reference type K.*

Now, consider two values of $\mathbf{z}$, say $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$. Differencing (5) across these two values:

$$\Lambda^{-1}\big(P(k,\mathbf{z}^{(1)})\big) - \Lambda^{-1}\big(P(k,\mathbf{z}^{(2)})\big) = b \times \big(\Delta_w^P(k,\mathbf{z}^{(1)}) - \Delta_w^P(k,\mathbf{z}^{(2)})\big) \tag{6}$$

or rearranging:

$$b = \frac{\Lambda^{-1}\big(P(k,\mathbf{z}^{(1)})\big) - \Lambda^{-1}\big(P(k,\mathbf{z}^{(2)})\big)}{\Delta_w^P(k,\mathbf{z}^{(1)}) - \Delta_w^P(k,\mathbf{z}^{(2)})} \tag{7}$$

From (7) it is clear that if the sign of the numerator and denominator are different, b is negative, otherwise b is positive. When are the signs different? Suppose that $\Delta_w^P(k,\mathbf{z}^{(1)}) - \Delta_w^P(k,\mathbf{z}^{(2)}) > 0$, so that in constituencies with characteristics $\mathbf{z}^{(1)}$ type k is relatively more likely to win than in constituencies with characteristics $\mathbf{z}^{(2)}$, and that $\Lambda^{-1}\big(P(k,\mathbf{z}^{(1)})\big) < \Lambda^{-1}\big(P(k,\mathbf{z}^{(2)})\big)$. Since $\Lambda^{-1}(\,\cdot\,)$ is increasing, this implies that

$$P(k,\mathbf{z}^{(1)}) < P(k,\mathbf{z}^{(2)})$$

or in words, that the party is less likely to select type k in constituencies with characteristics $\mathbf{z}^{(1)}$ than in constituencies with characteristics $\mathbf{z}^{(2)}$. So the parameter b is negative if the party tends to not select the candidate type that is relatively likely to win given the constituency characteristics $\mathbf{z}$.

With the parameter b identified, cost *differences* $c_k - c_K$ are identified as:

$$c_k - c_K = \Lambda^{-1}\big(P(k,\mathbf{z})\big) - b \times \big(w^P(k,\mathbf{z}) - w^P(K,\mathbf{z})\big) \tag{8}$$

and clearly, the difference $c_k - c_K$ is increasing in the choice probability $P(k,\mathbf{z})$, all else constant.

With this simpler case established, we now move to the case of type specific parameters b_k. Differencing Equation (4) across two values of $\mathbf{z}$ gives:

$$\Lambda^{-1}\big(P(k,\mathbf{z}^{(1)})\big) - \Lambda^{-1}\big(P(k,\mathbf{z}^{(2)})\big) = b_k \times \big(w^P(k,\mathbf{z}^{(1)}) - w^P(k,\mathbf{z}^{(2)})\big) - b_K \times \big(w^P(K,\mathbf{z}^{(1)}) - w^P(K,\mathbf{z}^{(2)})\big)$$

Now define:

$$\Delta_w^P(k,\mathbf{z}^{(1,2)}) \equiv w^P(k,\mathbf{z}^{(1)}) - w^P(k,\mathbf{z}^{(2)}), \quad k = 1, 2, ..., K \tag{9}$$

$$\Delta_\Lambda\big(k,\mathbf{z}^{(1,2)}\big) \equiv \Lambda^{-1}\big(P(k,\mathbf{z}^{(1)})\big) - \Lambda^{-1}\big(P(k,\mathbf{z}^{(2)})\big) \tag{10}$$

We can then re-write the difference in inverted choice probabilities as:

$$\Delta_\Lambda\big(k,\mathbf{z}^{(1,2)}\big) = b_k \times \Delta_w^P(k,\mathbf{z}^{(1,2)}) - b_K \times \Delta_w^P(K,\mathbf{z}^{(1,2)}) \tag{11}$$

and isolating for the reference parameter b_K we get:

$$b_K = \frac{b_k \times \Delta_w^P(k,\mathbf{z}^{(1,2)}) - \Delta_\Lambda\big(k,\mathbf{z}^{(1,2)}\big)}{\Delta_w^P(K,\mathbf{z}^{(1,2)})} \tag{12}$$

This holds at any pair of $\mathbf{z}$ vectors, so we can also write:

$$b_K = \frac{b_k \times \Delta_w^P(k,\mathbf{z}^{(2,3)}) - \Delta_\Lambda\big(k,\mathbf{z}^{(2,3)}\big)}{\Delta_w^P(K,\mathbf{z}^{(2,3)})} \tag{13}$$

and thus solve for the parameter b_k:

$$b_k = \frac{\Delta_w^P(K,\mathbf{z}^{(1,2)})\Delta_\Lambda\big(k,\mathbf{z}^{(2,3)}\big) - \Delta_w^P(K,\mathbf{z}^{(2,3)})\Delta_\Lambda\big(k,\mathbf{z}^{(1,2)}\big)}{\Delta_w^P(K,\mathbf{z}^{(1,2)})\Delta_w^P\big(k,\mathbf{z}^{(2,3)}\big) - \Delta_w^P(K,\mathbf{z}^{(2,3)})\Delta_w^P\big(k,\mathbf{z}^{(1,2)}\big)} \quad k = 1, ..., K-1 \tag{14}$$

Again, the parameter b_k is negative when the numerator and denominator have the opposite sign. When do they have the opposite sign? Suppose that $\Delta_w^P(K,\mathbf{z}^{(1,2)}) \simeq \Delta_w^P(K,\mathbf{z}^{(2,3)})$

so that Equation 14 reduces to

$$b_k = \frac{\Delta_\Lambda\big(k, \mathbf{z}^{(2,3)}\big) - \Delta_\Lambda\big(k, \mathbf{z}^{(1,2)}\big)}{\Delta_w^P\big(k, \mathbf{z}^{(2,3)}\big) - \Delta_w^P\big(k, \mathbf{z}^{(1,2)}\big)} \tag{15}$$

In this case, $b_k < 0$ if $\Delta_\Lambda\big(k, \mathbf{z}^{(2,3)}\big) < \Delta_\Lambda\big(k, \mathbf{z}^{(1,2)}\big)$ and $\Delta_w^P\big(k, \mathbf{z}^{(2,3)}\big) > \Delta_w^P\big(k, \mathbf{z}^{(1,2)}\big)$. Intuitively, this roughly can be interpreted to mean that the probability the party selects type k increases less moving from constituency $\mathbf{z}^{(3)}$ to constituency $\mathbf{z}^{(2)}$ than it does moving from constituency $\mathbf{z}^{(2)}$ to constituency $\mathbf{z}^{(1)}$ even though the probability of winning increases *more* moving from constituency $\mathbf{z}^{(3)}$ to constituency $\mathbf{z}^{(2)}$ than it does moving from constituency $\mathbf{z}^{(2)}$ to constituency $\mathbf{z}^{(1)}$.

The parameter on the reference type b_K is also identified by substituting Equation 14 into Equation 12, and cost differences are identified as:

$$c_k - c_K = \Lambda^{-1}\big(P(k, \mathbf{z})\big) - b_k \times w^P(k, \mathbf{z}) + b_K \times w^P(K, \mathbf{z}) \tag{16}$$

Note the following interesting features of the identification argument:

1. Variation in $\mathbf{z}$ is crucial for identifying b separately from c_k, and if we allow for type specific b we require more independent values of $\mathbf{z}$.

2. All type specific values of b are identified, but costs are only identified up to a reference type.

3 Model fit and validation

Here we provide further results about how model fit depends on the inclusion of party preferences over candidates.

In Table A.4 we present the analogue of Table 10 but in a model that assumes parties only care about voter preferences (and thus the probability of winning). This is the model estimated in the first column of the results Table 9.

Table A.4: Model fit with no cost parameters

Type	UPA actual	UPA predicted	NDA actual	NDA predicted	All actual	All predicted
1	217	130.49	229	147.47	446	277.96
2	24	95.66	43	87.89	67	183.55
3	49	97.68	22	92	71	189.68
4	144	110.17	140	106.64	284	216.81

When parties are restricted to care only about voter preference the model significantly under-predicts the selection of the educated type (type 1) and over-predicts the other types, in particular the uneducated type (type 2) and the Muslim type (type 3).