Online Appendix to "Do Parties and Voters Disagree? An Equilibrium Analysis of Candidate Selection in India"

Arvind Magesan
Economics Department, University of Calgary

Andrea Szabó
Economics Department, University of Houston

Gergely Ujhelyi
Economics Department, University of Houston

December 17, 2025

Contents

1 Data and sample

1.1 Construction of the Muslim characteristic

We construct a Muslim indicator based on candidates' names. First, we used the candidate affidavit data for all elections between 2008-2019 to obtain a library of Muslim names and common substrings ("name fragments") found in Muslim names. Indian Muslim names have Arabic or Farsi origins but are often spelled differently or modified from the original name in some other way. We create a similar library for non-Muslim names. There are a total of 470 names and fragments in the Muslim library and 1210 in the non-Muslim library.

For every name in our data which we wish to classify, we compute the "distance" between the name and 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. We take the distance of $name_i$ and a library (M or H) to be the minimum of these distances across names in the library. Let $d(name_i, M)$ and $d(name_i, H)$ denote these distances.

Next, we use the name fragments to construct another measure. We 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.

1.2 Sample construction

Our dataset is limited by the state constituencies for which we can obtain demographic information from the SHRUG (Asher et al. 2021). Because we analyze national elections, we also need to aggregate 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).[1]

[1]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).

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.[2]

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,[3] leaving us with a total of 232 constituencies in the dataset.

1.3 Missing characteristics

The specification of both voters' and parties' choices requires that all relevant candidate characteristics be observed. In the data, education, assets, and criminal history have missing values, and we impute these characteristics based on the candidate's gender, age, and SC/ST status (which have no missing values). Specifically, we impute assets, number of criminal cases, and number of completed years of education using the average by gender, SC/ST and age range (+/- 1 year relative to the candidate's age). For example, a 30 year old male non-SC/ST candidate's imputed asset is the average of all male non-SC/ST candidates aged 29-31. We then create the variables used in the estimation and the clustering algorithm (log(assets + 1), an indicator for at least one criminal case, and an indicator for at least twelve years of education). In practice, the main impact of imputing these characteristics is that we can use all candidates in the data when creating types.[4]

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, SC/ST and age) that we could use for imputing these missing values. The asset variable has 6,301 missing values and 562 zeros which we also treat as missing. Criminal history has 6,301, and education 7,629 missing values. The national election data has 6,581 observations. We drop 3 observations with missing characteristics that cannot be imputed. After aggregating small parties and independent candidates, we have 2,649 observations. Assets, criminal history, and education have, respectively, 119, 93 and 172 missing values. Among the 897 candidates of the UPA and the NDA, the corresponding numbers are 38, 36, and 53.

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

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

[4]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.

Table A.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				
Crime (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)	3425	0.73				
Muslim (0/1)	6581	0.12				
Crime (0/1)	3611	0.23				
Assets (log)	3544	14.63	2.51	14.76	11.46	17.60
Male (0/1)	6581	0.93				
Age	6581	46.25	11.96	45	31	63
SC or ST (0/1)	6581	0.35				
Incumbent (0/1)	6581	0.04				
Broadcast allowance (log)	6581	1.67	2.38	0.00	0.00	5.86
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 (share)	464	0.61	0.10	0.61	0.49	0.74
ST and SC population (share)	464	0.27	0.14	0.24	0.13	0.46
Rural population (share)	464	0.82	0.11	0.83	0.68	0.94
Population with paved roads (share)	464	0.83	0.19	0.90	0.58	1.00
Employed population (share)	464	0.42	0.07	0.43	0.32	0.50
Redistricting overlap (share)	464	0.87	0.17	1.00	0.59	1.00

Notes: Education: 1 if completed high school. Crime: 1 if has at least one criminal case. Assets (log): log of real assets in Rp. Incumbent: 1 if held the same seat just before the election. N. of candidates after aggregation: number of candidates after small-party candidates are aggregated. Redistricting overlap: largest share of a constituency's area in an old constituency.

1.4 Ideology measures

Wikipedia-based measures. We start from https://en.wikipedia.org/wiki/List_of_political_parties_in_India and the individual party websites linked from this page, and we collect information on the ideology of the 32 parties that enter our dataset as individual parties (i.e., are not in the "aggregate" party).

Table A.2: Coding of Left–Right ideology positions from Wikipedia

Description	Our code	N. parties	Description	Our code	N. parties
"far left"	-1	1	"center" or "syncretic"	0	8
"left wing to far left"	-0.75	1	"right wing" or "center right"	0.5	6
"left wing" or "center left"	-0.5	11	"right wing to far right"	0.75	1
"center to center left"	-0.25	3	"far right"	1	1
Total					32

Notes: Ideology categories from party descriptions on Wikipedia.

The first measure, Left-Right position, is coded as in Table A.2. If a party is in the NDA (respectively, UPA) alliance, we take its candidates' ideology to be the average of the party's and the BJP's (respectively, INC's) ideology. This reflects the idea that a party allied with one of these major parties is likely to campaign on a shared platform and present a blended ideological stance to voters.

The second measure uses Wikipedia's list of up to six specific ideological keywords for each party. Across the 32 parties in our sample, there are a total of 59 unique terms. We construct indicators for each ideological term, applying minimal grouping only when we are confident the terms refer to the same concept. If a party is in an alliance, we take the union of its ideology terms and those of the BJP (if NDA) or the INC (if UPA). After dropping perfectly co-linear variables, we are left with 12 unique ideological terms, shown in Table A.3.

Table A.3: Ideological keyword measures from Wikipedia

Term	N. parties	N. candidates	N. parties / alliances	Term	N. parties	N. candidates	N. parties / alliances
Regionalism	13	260	18	Integral	1	438	14
Agrarian	2	4	3	Islamic	1	2	1
Caste politics	9	121	11	Liberalism	1	455	12
Left	21	1574	29	Populism	5	308	7
Right	7	511	19	Progressivism	3	166	5
Dravidianism	3	42	4	Secularism	16	952	27

Notes: Keywords from Wikipedia party descriptions. N. parties/alliances shows the frequency of each term across the union of parties in an alliance.

Expert assessment based measures. We use information from the Democratic Accountability and Linkages Project (DALP; Kitschelt and Smidt (2013)), a dataset based on a 2009 survey of experts that scores parties along 5 dimensions (social spending on the disadvantaged, statism, public spending, national identity, traditional authority). These measures are only available for 15 of the 32 parties, and we assign values to the missing parties by predicting their scores based on the full set of Wikipedia terms described above.

First, we create an ideology index using principal components analysis (we find that the first principal component explains almost 80% of the variation in the data). Second, we focus on individual questions from the survey. In particular, Chhibber and Verma (2018) argue that there are two meaningful ideological dimensions among Indian parties, Recognition and Statism. We approximate the former with "Social spending on the disadvantaged," measured on a 1-10 scale where 1 is "Party advocates extensive social spending redistributing income to benefit the less well-off in society" and 10 is "Party opposes" such spending. To approximate Statism, we use "State role in governing the economy," measured on a 1-10 scale where 1 is "Party supports a major role for the state in regulating private economic activity to achieve social goals, in directing development, and/or maintaining control over key services." and 10 is "Party advocates a minimal role for the state in governing or directing economic activity or development." As above, when parties are in an alliance, we average their score with the BJP's (if NDA) or the INC's (if UPA).

2 Demand estimation: further details

2.1 Specification choice

Following Gandhi and Houde (2019), we first enter the differentiation IVs as controls in a Logit specification. This specification includes all the control variables, and instruments the endogenous characteristics as described in the paper. The results are in Table A.5. In column 1, the differentiation IVs for Muslim and Crime are statistically significant while the differentiation IVs for Education and Assets 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 Crime and obtain similar conclusions. Column 3 and 4 show corresponding estimates when Incumbent $\times$ Assets is included as another

Table A.4: Characteristics regressed on instruments

Dep. var.:	Education		Muslim		Crime		Assets		Incumbent		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)
IV^1	0.16	0.11	0.54	0.33	0.30	0.23	0.30	0.30	0.02	0.01	0.01
	(0.05)	(0.05)	(0.07)	(0.08)	(0.06)	(0.06)	(0.05)	(0.05)	(0.00)	(0.00)	(0.00)
IV^2		0.18		0.38		0.21		0.02		0.01	0.01
		(0.05)		(0.08)		(0.06)		(0.05)		(0.00)	(0.00)
IV^3											0.14
											(0.05)
Adj. R^2	0.15	0.16	0.09	0.10	0.14	0.14	0.36	0.36	0.15	0.16	0.16
F	10.07	11.14	61.04	45.21	23.78	17.44	41.95	20.27	83.02	48.32	39.39
Mean	0.31	0.31	-0.03	-0.03	0.08	0.08	-0.02	-0.02	0.10	0.10	0.10
Std. dev.	0.81	0.81	0.88	0.88	0.98	0.98	0.96	0.96	0.30	0.30	0.30

Notes: For Education, Muslim, Crime and Assets, IV^1 is the average of the characteristic among a party's candidates in state constituencies in the given year, and IV^2 is the same variable for the other election in the data (2009 for 2014 and vice versa). For Incumbent, IV^1 is the party's competitiveness in the previous election, $IV^2 = IV^1 \times \mathbf{I}(year = 2014)$, and IV^3 is redistricting overlap. Regressions include broadcast allowance, and fixed effects for state, year, party, alliance, imputed characteristics, and reserved constituencies. F is the Olea and Pflueger (2013) effective F statistic. The last rows show the mean and standard deviation of the dependent variable. Robust standard errors in parentheses. N = 2,649.

(endogenous) characteristic.

2.2 Sensitivity of the demand estimates

We follow the method proposed by Andrews et al. (2017) to evaluate the extent to which our demand estimates are sensitive to failures of the moment conditions, i.e., the exogeneity of our instruments. We use their measure to compute the bias in our main estimates assuming a violation in each moment condition where a 1 standard deviation change in the instrument results in a 1 unit (approximately one third standard deviation) change in the econometric error ξ. We compute this bias for each instrument used to identify the endogenous characteristics, and display the impact on 8 of our key parameters in Figure A.2. The vertical lines show the parameter estimates (column 1 of Table 1 in the paper) and the dots the "true" parameter once the implied bias is taken into account. Entries on the Y-axis correspond to the 11 instruments (see section 5.2.2 in the paper for details).

The figure shows that none of our parameter estimates is very sensitive to this kind of mis-specification. For example, our point estimate for the marginal utility of Assets, 2.8485, would be biased away from values between 2.848 and 2.851 if the exogeneity assumption on the corresponding instrument failed in the manner described above.

Table A.5: Specification choice: differentiation IVs

	(1)	(2)	(3)	(4)
Incumbent	3.15	3.19	3.81	3.78
	(0.97)	(0.96)	(1.22)	(1.20)
Assets	1.51	1.54	2.36	2.40
	(0.39)	(0.41)	(0.60)	(0.64)
Incumbent × Assets			-3.36	-3.06
			(1.37)	(1.29)
Education	-0.46	-0.52	-1.54	-1.57
	(0.45)	(0.48)	(0.70)	(0.75)
Muslim	-0.39	-0.41	-0.41	-0.43
	(0.18)	(0.18)	(0.23)	(0.24)
Crime	0.37	0.28	0.04	-0.04
	(0.30)	(0.30)	(0.39)	(0.39)
diffIV(Educ)	-0.02		-0.01	
	(0.03)		(0.04)	
diffIV(Muslim)	-0.07	-0.08	-0.05	-0.07
	(0.03)	(0.03)	(0.03)	(0.03)
diffIV(Crime)	0.07	0.06	0.07	0.03
	(0.03)	(0.03)	(0.04)	(0.05)
diffIV(Assets)	-0.03		-0.10	
	(0.04)		(0.06)	
J p-val	0.00	0.00	0.15	0.14
IIA p-val	0.00	0.00	0.00	0.06

Notes: Gandhi and Houde (2019) specification checks. The dependent variable is vote shares. Candidate characteristics are instrumented as described in section 5.2.2 in the paper, and the differentiation IVs are entered as controls. Specifications also control for broadcast allowance and fixed effects for state, year, party, alliance, 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 A.6: Specification choice: differentiation IVs and voter demographics

	Lit	Rural	Roads	Empl	SC/ST	Lit	Rural	Roads	Empl	SC/ST	SC/ST / Rural
Crime	0.02	0.05	-0.00	0.13	0.18						0.18
	(0.07)	(0.05)	(0.05)	(0.10)	(0.10)						(0.10)
Muslim						-0.12	-0.07	-0.10	-0.13	-0.05	-0.09
						(0.05)	(0.04)	(0.03)	(0.07)	(0.07)	(0.03)
J p-val	0.15	0.14	0.18	0.12	0.14	0.14	0.13	0.16	0.14	0.13	0.14
IIA p-val	0.06	0.06	0.06	0.05	0.04	0.10	0.07	0.02	0.12	0.24	0.05

Notes: Gandhi and Houde (2019) specification checks. The dependent variable is vote shares. Candidate characteristics are instrumented as described in section 5.2.2 in the paper, and the differentiation IVs are entered as controls. The first 5 columns include diffIVs for Crime, the next five include diffIVs for Muslim. The last column includes the diffIV for SC/ST × Crime and Muslim × Rural. Only the coefficients on the diffIVs are shown. Specifications control for broadcast allowance and fixed effects for state, year, party, alliance, 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.

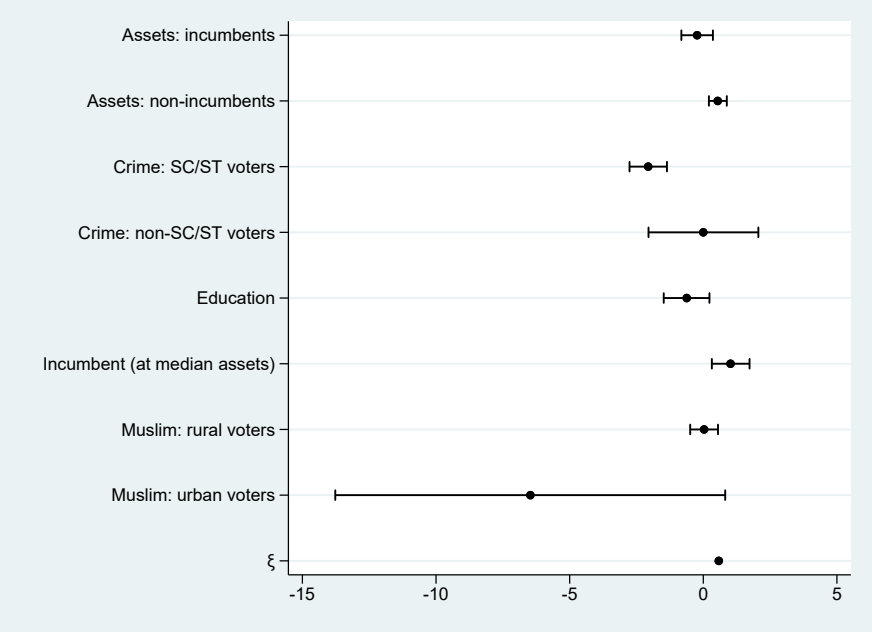

Figure A.1: Magnitude of voter preference for different candidate characteristics

Notes: Estimated effect of increasing the listed characteristic from its lowest to its highest value (if binary) or by 1 standard deviation (if continuous) on voter utility, measured in standard deviation units. The effect of Asset is shown separately for incumbent and non-incumbent candidates, and the effect of Muslim and Crime is shown separately for different voter groups. Values based on the specification in Table 1, column 1. Bands show the 95% confidence intervals.

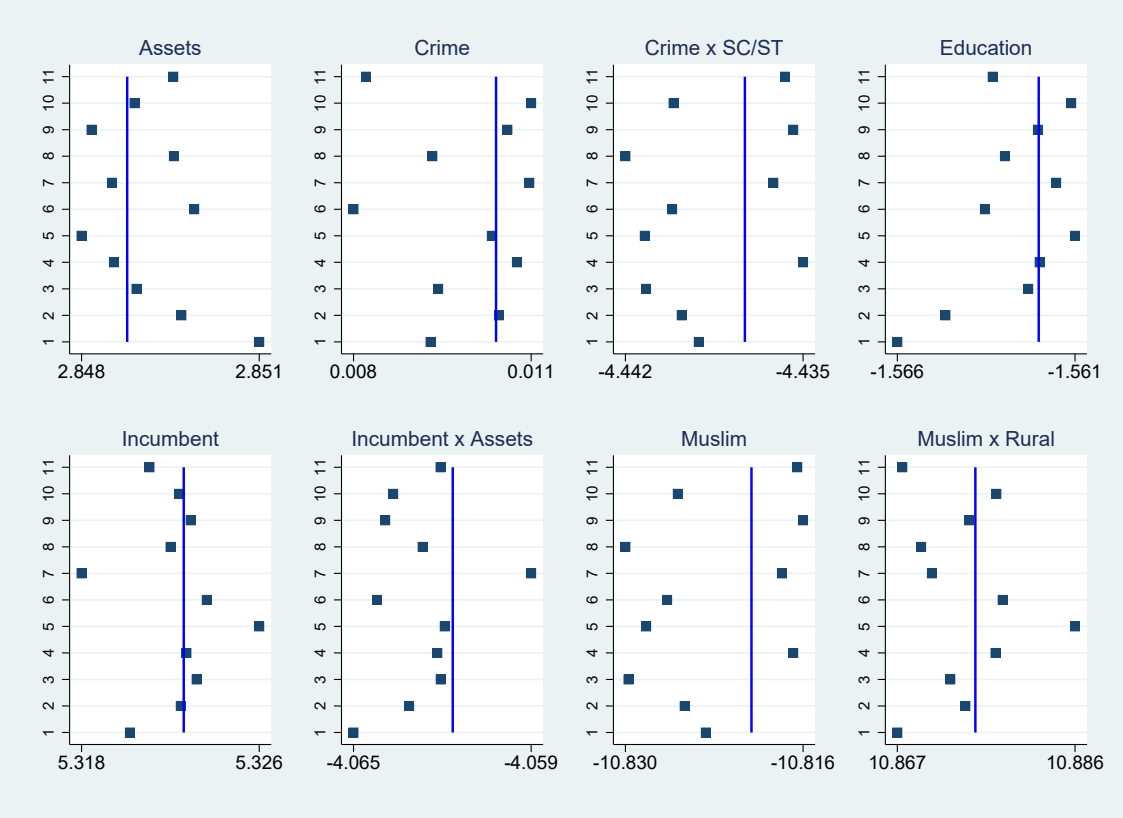

Figure A.2: Sensitivity of demand estimates to failures of the moment conditions

Notes: Andrews et al. (2017) sensitivity measures for the demand parameters listed. On each panel, the vertical line shows the point estimate in our main specification. The dots show the "true" parameter (our estimate minus its bias) assuming that a 1 standard deviation change in one of the instruments results in a 1 unit change in the econometric error ξ. The instruments listed on the Y axis are 1-2: Education instruments, 3-4: Muslim instruments, 5-6: Crime instruments, 7: Asset instrument, 8-10: Incumbent instruments, 11: Instrument for Asset $\times$ Incumbent.

Table A.7: Additional demand specifications

	(1)	(2)	(3)	(4)	(5)
Incumbent	5.46	4.46	4.12	5.60	8.74
	(1.92)	(1.96)	(1.74)	(1.80)	(4.39)
Assets	2.67	2.63	2.80	2.79	5.22
	(0.81)	(1.01)	(0.91)	(0.97)	(3.93)
Incumbent × Assets	-3.82	-3.6	-3.77	-4.02	-8.26
	(1.60)	(1.48)	(1.55)	(1.76)	(6.91)
Education	-1.32	-1.46	-1.67	-1.61	-1.47
	(0.96)	(0.96)	(0.84)	(1.35)	(1.70)
Muslim	-9.67	0.26	0.35	-11.49	-11.17
	(5.69)	(2.72)	(3.45)	(6.14)	(8.92)
Crime	0.17	0.41	0.34	-0.06	0.66
	(0.90)	(0.49)	(0.92)	(0.73)	(0.93)
Nonlinear parameters:					
Crime × SC/ST	-5.28	-4.47	-3.43	-4.56	-9.72
	(3.02)	(2.50)	(4.13)	(2.52)	(6.64)
Muslim × Rural	9.67	-0.86	-0.94	11.50	11.14
	(5.70)	(3.97)	(4.58)	(6.21)	(8.97)
v_{Crime}	-0.11		-0.17		
	(7.15)		(6.84)		
v_{Muslim}		0.06	0.02		
		(10.95)	(16.59)		
$v_{Education}$				0.07	
				(14.55)	
v_{Asset}					-2.91
					(2.53)
J	5.42	10.51	10.63	4.68	2.11
p-value	0.37	0.06	0.06	0.46	0.83
Newey-West pval	0.12	0.34	0.55	0.16	0.41

Notes: BLP estimates. Specifications include broadcast allowance and fixed effects for state, year, party, alliance, imputed characteristics, and reserved constituencies. J is the overidentification J-statistic (df=5) with its p-value. The bottom row shows the p-value of the Newey-West D-test for the null that all nonlinear parameters are jointly 0. Robust standard errors clustered by constituency in parentheses. N = 2,649.

3 Identification of model parameters

We formally establish identification of party payoffs and discuss the intuition of the identification results. As is standard in the empirical game literature (see, for example, Bajari et al. (2013)), we assume that choice probabilities are known to the researcher.[5] Then given vote shares, win probabilities and expected vote shares are also known and respectively given by $w_p^P(k,\mathbf{z}) = E_p[w_p(a_p, a_{-p}, \mathbf{z})|a_p = k]$ and $s_p^P(k,\mathbf{z}) = E_p[s_p(a_p, a_{-p}, \mathbf{z})|a_p = k]$ for choice $a_p = k$, $k = 1, ..., K$, where the expectation $E_p[w_p(a_p, a_{-p}, \mathbf{z})|a_p = k]$ is an integration over a_{-p} using player $-p$'s choice probability and $\mathbf{z}$ is the vector of constituency specific observable payoff variables.

We begin with a simple case with no costs so the only parameters to identify are (b^w, b^s). With identification in this simple case established, we then re-introduce cost parameters below. In this simple case, party p's choice probability satisfies

$$P_p(k,\mathbf{z}) = \Lambda\big(b^w \times w_p^P(k,\mathbf{z}) + b^s \times s_p^P(k,\mathbf{z})\big) = \frac{\exp\left\{b^w \times w_p^P(k,\mathbf{z}) + b^s \times s_p^P(k,\mathbf{z})\right\}}{\sum_{k'} \exp\left\{b^w \times w_p^P(k',\mathbf{z}) + b^s \times s_p^P(k',\mathbf{z})\right\}}$$

(As the argument for identification is symmetric across players, we drop the p subscript in what follows.) Since $\Lambda(\cdot)$ is strictly increasing, inverting the choice probability gives:

$$\begin{aligned} \Lambda^{-1}\big(P(k,\mathbf{z})\big) &= \ln\big(P(k,\mathbf{z})\big) - \ln\big(P(K,\mathbf{z})\big) \\ &= b^w \times (w^P(k,\mathbf{z}) - w^P(K,\mathbf{z})) + b^s \times (s^P(k,\mathbf{z}) - s^P(K,\mathbf{z})) \end{aligned} \tag{1}$$

where we have taken type K as the reference type. Define $\Delta_w^P(k,\mathbf{z}) \equiv w^P(k,\mathbf{z}) - w^P(K,\mathbf{z})$ and $\Delta_s^P(k,\mathbf{z}) \equiv s^P(k,\mathbf{z}) - s^P(K,\mathbf{z})$ and consider two constituencies, $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$. These satisfy a system of equations:

$$\begin{bmatrix} \Lambda^{-1}\big(P(k,\mathbf{z}^{(1)})\big) \\ \Lambda^{-1}\big(P(k,\mathbf{z}^{(2)})\big) \end{bmatrix} = \begin{bmatrix} \Delta_w^P(k,\mathbf{z}^{(1)}) & \Delta_s^P(k,\mathbf{z}^{(1)}) \\ \Delta_w^P(k,\mathbf{z}^{(2)}) & \Delta_s^P(k,\mathbf{z}^{(2)}) \end{bmatrix} \begin{bmatrix} b^w \\ b^s \end{bmatrix} \tag{2}$$

and the parameters (b^w, b^s) are identified under the rank condition:

$$\frac{\Delta_w^P(k,\mathbf{z}^{(1)})}{\Delta_w^P(k,\mathbf{z}^{(2)})} \neq \frac{\Delta_s^P(k,\mathbf{z}^{(1)})}{\Delta_s^P(k,\mathbf{z}^{(2)})} \tag{3}$$

Roughly, $\mathbf{z}$ should shift expected vote share independently of win probability. As vote shares can change without turning a loss into a win, this condition is easy to satisfy and b^w, b^s are identified.

[5] Consistent estimates of choice probabilities can be obtained, in principle, using a raw frequency estimator.

One of our key findings is that $b^w > 0 > b^s$. We illustrate how this can arise as well as the intuition given the identification argument. Solving the system in (2) we get

$$b^w = \frac{\Delta_s^P(k, \mathbf{z}^{(2)})\Lambda^{-1}(P(k, \mathbf{z}^{(1)})) - \Delta_s^P(k, \mathbf{z}^{(1)})\Lambda^{-1}(P(k, \mathbf{z}^{(2)}))}{\Delta_s^P(k, \mathbf{z}^{(2)})\Delta_w^P(k, \mathbf{z}^{(1)}) - \Delta_s^P(k, \mathbf{z}^{(1)})\Delta_w^P(k, \mathbf{z}^{(2)})} \tag{4}$$

and

$$b^s = \frac{\Delta_w^P(k, \mathbf{z}^{(1)})\Lambda^{-1}(P(k, \mathbf{z}^{(2)})) - \Delta_w^P(k, \mathbf{z}^{(2)})\Lambda^{-1}(P(k, \mathbf{z}^{(1)}))}{\Delta_s^P(k, \mathbf{z}^{(2)})\Delta_w^P(k, \mathbf{z}^{(1)}) - \Delta_s^P(k, \mathbf{z}^{(1)})\Delta_w^P(k, \mathbf{z}^{(2)})} \tag{5}$$

Assume without loss that the denominator of Equations (4) and (5) is positive. This would occur if, for example, $\Delta_s^P(k, \mathbf{z}^{(2)}) > \Delta_s^P(k, \mathbf{z}^{(1)})$ and $\Delta_w^P(k, \mathbf{z}^{(2)}) < \Delta_w^P(k, \mathbf{z}^{(1)})$: Type k delivers relatively higher expected vote share but lower win probability in constituency 2 relative to constituency 1. For $b^w > 0$ in this case, it is sufficient that: $P(k, \mathbf{z}^{(1)}) > P(k, \mathbf{z}^{(2)})$. In other words, $b^w > 0$ rationalizes type k being chosen more frequently in constituency 1 than in constituency 2 even though $\Delta_s^P(k, \mathbf{z}^{(2)}) > \Delta_s^P(k, \mathbf{z}^{(1)})$.

Turning to Equation (5), for $b^s < 0$ given that $\Delta_w^P(k, \mathbf{z}^{(2)}) < \Delta_w^P(k, \mathbf{z}^{(1)})$, $P(k, \mathbf{z}^{(1)})$ must be *sufficiently* larger than $P(k, \mathbf{z}^{(2)})$. So in sum, if $\Delta_s^P(k, \mathbf{z}^{(2)}) > \Delta_s^P(k, \mathbf{z}^{(1)})$ and $\Delta_w^P(k, \mathbf{z}^{(2)}) < \Delta_w^P(k, \mathbf{z}^{(1)})$ then $P(k, \mathbf{z}^{(1)}) >> P(k, \mathbf{z}^{(2)})$ is rationalized by $b^w > 0$, $b^s < 0$.

Adding costs to the payoff function is straightforward. In this case the inverted choice probability satisfies:

$$\begin{aligned} \Lambda^{-1}\big(P(k, \mathbf{z})\big) &= \ln\big(P(k, \mathbf{z})\big) - \ln\big(P(K, \mathbf{z})\big) \\ &= b^w \times (w^P(k, \mathbf{z}) - w^P(K, \mathbf{z})) + b^s \times (s^P(k, \mathbf{z}) - s^P(K, \mathbf{z})) - (c_k - c_K) \end{aligned} \tag{6}$$

Now, consider two values of $\mathbf{z}$, say $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(2)}$. Differencing (6) across these two values we eliminate the cost parameters, assumed to be independent of $\mathbf{z}$. We can do the same at another pair of value of $\mathbf{z}$, say $\mathbf{z}^{(1)}$ and $\mathbf{z}^{(3)}$, which gives us a system of two equations and two unknown parameters (b^w, b^s):

$$\begin{bmatrix} \Delta_\Lambda^P(k, \mathbf{z}^{(1,2)}) \\ \Delta_\Lambda^P(k, \mathbf{z}^{(1,3)}) \end{bmatrix} = \begin{bmatrix} \Delta_w^P(k, \mathbf{z}^{(1,2)}) & \Delta_s^P(k, \mathbf{z}^{(1,2)}) \\ \Delta_w^P(k, \mathbf{z}^{(1,3)}) & \Delta_s^P(k, \mathbf{z}^{(1,3)}) \end{bmatrix} \begin{bmatrix} b^w \\ b^s \end{bmatrix} \tag{7}$$

where $\Delta_\Lambda^P(k, \mathbf{z}^{(1,2)}) \equiv \Lambda^{-1}\big(P(k, \mathbf{z}^{(1)})\big) - \Lambda^{-1}\big(P(k, \mathbf{z}^{(2)})\big)$ (and similarly for $\Delta_\Lambda^P(k, \mathbf{z}^{(1,3)})$), and $\Delta_w^P(k, \mathbf{z}^{(1,2)}) \equiv \Delta_w^P\big(P(k, \mathbf{z}^{(1)})\big) - \Delta_w^P\big(P(k, \mathbf{z}^{(2)})\big)$ (and similarly for $\Delta_s^P(k, \mathbf{z}^{(1,2)})$).

The parameters (b^w, b^s) are identified if, for at least one triple $(\mathbf{z}^1, \mathbf{z}^2, \mathbf{z}^3)$:

$$\frac{\Delta_w^P(k, \mathbf{z}^{(1,2)})}{\Delta_w^P(k, \mathbf{z}^{(1,3)})} \neq \frac{\Delta_s^P(k, \mathbf{z}^{(1,2)})}{\Delta_s^P(k, \mathbf{z}^{(1,3)})}$$

This condition is analogous to the condition in Equation (3) in the simpler case, except we now need variation in win probability and expected vote share across more values of $\mathbf{z}$ - this is how we identify b^w, b^s separately from the cost difference $c_k - c_K$.

Solving the system in (7) we get expressions for b^w and b^s analogous to equations (4) and (5) in the simple case with no costs. With b^w and b^s identified we can obtain an expression for the cost difference $c_k - c_K$ by substituting the identified values into (6).

4 Clustering and candidate types

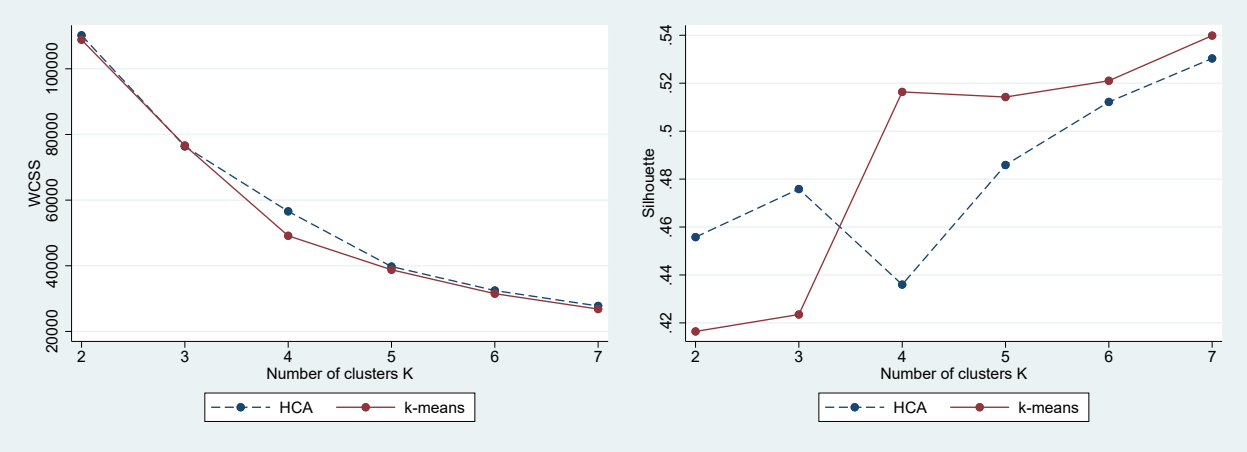

Figure A.3: WCSS and Silhouette

Notes: The Within-Cluster Sum of Squares (WCSS) measures the similarity of units within clusters. The Silhouette coefficient measures the dissimilarity of units across clusters. The graph shows the values of these measures that result from running the k-means or HCA clustering algorithm for different numbers of clusters K in our data.

We illustrate that standard methods of validation for unsupervised clustering methods support $K = 4$ using k-means in our application and compare our k-means approach to an alternative clustering algorithm, Hierarchical Clustering Analysis (HCA).

HCA begins by treating each of the N individual observation as its own "cluster," then iteratively agglomerates the clusters. In the first iteration, we take the two points that are closest in the sense that they have the lowest variance among any pair of points in the sample (the Ward merging criterion), and then combine those to form a new cluster. We now have $N-1$ clusters. We again find the two clusters that are closest and combine them, and so on. The number of clusters is determined by where we stop this iterative agglomeration process.

On the left panel of Figure A.3, we compare the within-cluster sum of squares, WCSS, over a number of possible clusters across the two methods. The k-means approach is uniformly better by this metric (i.e., produces lower WCSS scores).[6] On the right panel, we compare the Silhouette score associated with the two approaches. While HCA performs better for 2 or 3 clusters, k-means performs better from 4 clusters on. Moreover, k-means scores with 4 or more clusters dominate any HCA score below 4 clusters. Finally, note that we only ever hit a Silhouette score of 0.5 with k-means and the number of clusters at least 4. (In cluster analysis, a Silhouette score of 0.5 is often used as a rule of thumb for "good clustering" (Rousseeuw 1987).) These patterns indicate that k-means is preferable to HCA for our application.

Nevertheless, we computed the types that the HCA algorithm predicts when we set $k = 4$. These types look remarkably similar to what we obtain using k-means. Similar to k-means, HCA clusters all Muslim candidates together, and then among the remaining candidates clusters all criminals together. The one difference is that the HCA algorithm puts some uneducated individuals into Type 1 so that there is a larger gap in assets between the Type 1 and Type 2 candidates than there is using k-means.

Table A.8: Distribution of candidate types

	State elections			National elections			
	All candidates	UPA	NDA	All candidates	UPA	NDA	Incumbents
Type 1	34.74	46.78	44.15	37.59	50.00	52.53	47.03
Type 2	39.63	16.37	20.76	28.78	5.53	9.45	7.03
Type 3	11.20	11.27	6.55	12.71	11.29	5.07	8.11
Type 4	14.43	25.58	28.55	20.91	33.18	32.95	37.84
Total	100	100	100	100	100	100	100
N	36,764	3,788	4,109	6,049	434	434	185

Notes: Percentage of candidates in state and national elections assigned to each type by the k-means clustering algorithm. Types 1-4 are the educated, uneducated, Muslim, and criminal types, respectively. National elections are the constituencies in the final estimation sample. The last column shows the distribution of re-nominated UPA or NDA incumbents.

5 Additional results with ideology measures

[6]Note that k-means is based on centroids (the average of all points in the cluster), while HCA is not. To generate Figure A.3, we compute the centroid of each HCA cluster analogously.

Table A.9: Voter preference parameter estimates with heterogeneity on parties

	(1)	(2)	(3)
Incumbent	3.46	6.81	3.12
	(2.23)	(2.88)	(1.82)
Assets	2.70	2.53	2.85
	(0.68)	(0.81)	(1.43)
Incumbent × Assets	-3.48	-4.19	-3.56
	(2.59)	(1.88)	(1.53)
Education	-0.57	-1.39	-1.30
	(0.78)	(1.09)	(1.25)
Muslim	0.04	-8.46	0.11
	(4.90)	(5.59)	(3.54)
Crime	0.79	0.25	0.05
	(0.48)	(0.83)	(0.85)
Ideology (Left-Right)	1.24	2.84	1.06
	(1.70)	(1.50)	(1.11)
Nonlinear parameters:			
Muslim × Rural	-0.55	8.13	-0.82
	(6.13)	(5.67)	(4.83)
Crime × SC/ST	-4.08	-5.41	0.81
	(5.48)	(2.49)	(4.08)
v_{UPA}	-0.04		
	(57.05)		
v_{NDA}	-0.01		
	(38.88)		
UPA × Roads		-5.50	
		(2.69)	
NDA × Roads		13.53	
		(21.47)	
UPA × SC/ST			-5.38
			(24.10)
NDA × SC/ST			-0.08
			(5.31)
J	18.00	6.32	14.64
p-value	0.00	0.28	0.01
Newey-West pval	0.59	0.05	0.75

Notes: BLP estimates. Column 1 includes Normally distributed random coefficients on the UPA and NDA dummies. Columns (2) and (3) allow for heterogeneity on these dummies by access to paved roads and SC/ST.

Table A.10: Additional voter preference parameter estimates with alternative ideology measures

	Left-Right	Left-Right	Left-Right	DALP	DALP	DALP
Incumbent	3.68	3.86	4.85	5.01	4.57	5.19
	(1.71)	(2.08)	(1.87)	(1.71)	(1.77)	(1.92)
Assets	3.39	3.69	3.17	2.05	2.88	2.73
	(1.11)	(0.99)	(1.14)	(0.56)	(1.24)	(0.88)
Incumbent × Assets	-4.42	-5.06	-4.31	-2.68	-4.05	-3.31
	(1.72)	(1.93)	(2.51)	(1.22)	(2.00)	(1.77)
Education	-1.54	-1.79	-1.53	0.18	-0.80	-1.27
	(1.23)	(1.19)	(1.00)	(0.72)	(1.33)	(1.18)
Muslim	-7.25	-11.84	-12.23	-6.92	-9.96	-11.90
	(5.85)	(7.38)	(6.89)	(4.62)	(6.46)	(6.84)
Crime	0.57	0.01	0.23	0.89	0.47	0.51
	(0.65)	(0.83)	(0.87)	(0.45)	(0.67)	(0.72)
Ideology	2.46	-0.13	1.84	0.50	0.14	0.46
	(1.55)	(2.45)	(2.87)	(0.20)	(0.58)	(0.27)
Nonlinear parameters:						
Muslim × Rural	7.25	11.94	12.28	6.91	9.98	11.91
	(5.96)	(7.46)	(6.92)	(4.67)	(6.51)	(6.82)
Crime × SC/ST	-4.69	-3.06	-4.69	-6.52	-5.55	-4.84
	(2.64)	(3.67)	(2.58)	(1.85)	(2.41)	(2.77)
v_{Ideology}	0.02			-0.01		
	(34.03)			(4.48)		
Ideology × Roads		3.44			0.69	
		(2.19)			(0.80)	
Ideology × SC/ST			0.84			0.51
			(4.55)			(0.32)
J	7.82	5.07	3.78	11.86	7.81	4.96
p-value	0.17	0.41	0.58	0.04	0.17	0.42
Newey-West pval	0.53	0.18	0.16	0.06	0.07	0.05

Notes: BLP estimates. Column headers show the ideology measures used. In each case, the first specification includes a Normally distributed random coefficient on Ideology, the second allows for heterogeneity by voters' access to paved roads, and the third allows for heterogeneity by voter's SC/ST status.

Table A.11: Party objective function estimates with ideology measure

	(1)	(2)		(1)	(2)
b^w	3.18	2.75	c_4^{educ}	0.46	0.56
	(0.77)	(0.85)		(0.42)	(0.40)
b^s	-8.52	-6.70	c_1^{crime}	0.92	0.98
	(1.13)	(1.17)		(0.28)	(0.27)
$c_{1,NDA}^0$	2.48	1.78	c_2^{crime}	0.77	0.82
	(0.50)	(0.45)		(0.36)	(0.36)
$c_{2,NDA}^0$	3.36	2.74	c_3^{crime}	0.46	0.53
	(0.50)	(0.45)		(0.36)	(0.35)
$c_{3,NDA}^0$	4.49	3.78	c_4^{crime}	-0.13	-0.05
	(0.56)	(0.51)		(0.29)	(0.28)
$c_{4,NDA}^0$	3.55	2.78	c_1^{asset}	0.40	0.19
	(0.50)	(0.45)		(0.33)	(0.31)
$c_{1,UPA}^0$	1.91	1.48	c_2^{asset}	0.36	0.09
	(0.41)	(0.38)		(0.41)	(0.39)
$c_{2,UPA}^0$	3.50	3.10	c_3^{asset}	0.83	0.61
	(0.49)	(0.48)		(0.41)	(0.40)
$c_{3,UPA}^0$	3.22	2.77	c_4^{asset}	0.03	-0.09
	(0.48)	(0.45)		(0.34)	(0.32)
$c_{4,UPA}^0$	2.91	-2.38	c_1^{Muslim}	0.28	0.32
	(0.43)	(0.39)		(0.40)	(0.39)
c_1^{educ}	0.28	0.33	c_2^{Muslim}	0.26	0.30
	(0.40)	(0.38)		(0.56)	(0.56)
c_2^{educ}	0.77	0.85	c_3^{Muslim}	-1.05	-1.00
	(0.51)	(0.50)		(0.45)	(0.43)
c_3^{educ}	0.54	0.58	c_4^{Muslim}	0.05	0.14
	(0.52)	(0.51)		(0.43)	(0.41)
Log likelihood				-819.31	-836.41

Notes: Column 1 uses the demand specification in column 1 of Table 6 in the paper. Column 2 uses the demand specification in column 2 of Table A.10. The cost parameters c are measured relative to the excluded category, incumbents. $c_{k,p}^0$ is party p's direct cost of choosing a type k candidate. c_k^l is the impact of characteristic l in the pool of candidates on the recruitment cost of a type k candidate. Types 1-4 are the educated, uneducated, Muslim, and criminal types, respectively. Standard errors in parentheses. Number of markets: 434.

6 Criminals as strategic complements

To obtain a precise measure of strategic complementarity, write party p's choice probability associated with choosing some type a' as $P_p(a') = \frac{\exp\left\{\tilde{U}_p(a',P)\right\}}{\sum_{a=1}^{K}\exp\left\{\tilde{U}_p(a,P)\right\}}$, where $\tilde{U}_p(a,P) = E_P[b^w w_p(a)+b^s s_p(a)]+c_p(a)$ and K is 5 if the party can choose to retain the incumbent and 4 otherwise. Letting $\Delta^P(a) \equiv \tilde{U}_p(a,P)-\tilde{U}_p(4,P)$, we have $P_p(4) = 1/\left(1+\sum_{a\neq 4}\exp\left\{\Delta^P(a)\right\}\right)$. Differentiating with respect to the opponent's probability of choosing type 4, we get

$$\frac{\partial P_p(4)}{\partial P_{-p}(4)} = -\sum_{a\neq 4}\frac{\exp\left\{\Delta^P(a)\right\}\frac{\partial \Delta^P(a)}{\partial P_{-p}(4)}}{\left(1+\sum_{a\neq 4}\exp\left\{\Delta^P(a)\right\}\right)^2} = -\sum_{a\neq 4} P_p(4)P_p(a)\frac{\partial \Delta^P(a)}{\partial P_{-p}(4)}$$

We have

$$\frac{\partial \Delta^P(a)}{\partial P_{-p}(4)} = b^w\frac{\partial E_P[w_p(a)-w_p(4)]}{\partial P_{-p}(4)} + b^s\frac{\partial E_P[s_p(a)-s_p(4)]}{\partial P_{-p}(4)}.$$

where

$$\frac{\partial E_P[w_p(a)-w_p(4)]}{\partial P_{-p}(4)} = w_p(a,4)-w_p(4,4)+\sum_{t\neq 4}\left[w_p(a,t)-w_p(4,t)\right]\frac{\partial P_{-p}(t)}{\partial P_{-p}(4)}, \tag{8}$$

and similarly for $\frac{\partial E_P[s_p(a)-s_p(4)]}{\partial P_{-p}(4)}$.

Using our parameter estimates, we compute (8) for both parties in every constituency. We do this by assuming that $\frac{\partial P_{-p}(t)}{\partial P_{-p}(4)}$ in (8) is the same for all $t \neq 4$ (if we instead set $\frac{\partial P_{-p}(t)}{\partial P_{-p}(4)} = -1$ for some t we obtain very similar results). A positive derivative $\frac{\partial P_p(4)}{\partial P_{-p}(4)}$ indicates strategic complementarity between criminal candidates.

Among constituencies with no incumbent available, we find that the derivative is positive for both parties in 84%, and positive for at least one of the parties in 99%.

7 Additional robustness checks

7.1 Alternative payoff specifications

Table A.12 probes the robustness of our results, in particular $b^s < 0$, to modifications of parties' payoffs. The first two columns include nonlinearities in win probability, and we continue to find $b^s < 0$. Columns 3 and 4 replace vote share with vote margin or number of votes, respectively, as alternative measures of popularity. Both of these measures also yield a significant negative coefficient.

In Table A.13, we add an (expected) interaction term

$$\sum_{\mathbf{a}_{-p}} w_p(a_p, \mathbf{a}_{-p}) s_p(a_p, \mathbf{a}_{-p}) P(\mathbf{a}_{-p})$$

to the party payoffs. This allows for the possibility that parties value expected vote share differentially, depending on whether the share is associated with an anticipated win or a loss. According to the estimates, parties have a negative but insignificant payoff from more popular candidates who are likely to lose, and a significant negative payoff from candidates who are likely to win. This confirms that our main estimates in the paper are driven by parties' aversion to candidates who are expected to win big.

Table A.12: Alternative payoff specifications

	(1)	(2)	(3)	(4)
b^w	8.29	4.87	3.00	3.85
	(1.20)	(2.29)	(0.75)	(0.80)
b_2^w	-5.68	3.54		
	(1.17)	(5.33)		
b_3^w		-6.78		
		(3.85)		
b^s	-8.53	-8.40		
	(1.34)	(1.36)		
b^m			-6.65	
			(0.81)	
b^n				-0.73
				(0.08)
Log likelihood	-782.00	-779.93	-809.89	-796.14

Notes: Estimates of alternative payoff specifications. Cost parameters c are not shown. Column 1 adds squared win probability (the expected value of w^2) and column 2 also adds cubed win probability. Column 3 replaces expected vote share s with expected vote margin (with a corresponding parameter b^m) and column 4 with expected number of votes (with parameter b^n). Standard errors in parentheses. Number of markets: 434.

Table A.13: Party objective functions, with interaction between win probability and vote share

	(1)	(2)	(3)		(1)	(2)	(3)
b^w	7.50	4.79	4.30	c_4^{educ}		2.03	0.26
	(0.68)	(0.86)	(0.90)			(0.41)	(0.49)
b^s	-12.70	-2.98	-2.09	c_1^{crime}		1.71	0.95
	(1.58)	(1.94)	(2.20)			(0.33)	(0.34)
$b^{w,s}$	2.01	-7.51	-8.79	c_2^{crime}		1.82	0.74
	(1.29)	(1.59)	(1.81)			(0.38)	(0.41)
$c_{1,NDA}^0$			3.02	c_3^{crime}		1.72	0.48
			(0.60)			(0.37)	(0.41
$c_{2,NDA}^0$			4.06	c_4^{crime}		0.94	-0.10
			(0.60)			(0.33)	(0.35)
$c_{3,NDA}^0$			5.09	c_1^{asset}		2.04	0.65
			(0.64)			(0.33)	(0.39)
$c_{4,NDA}^0$			4.14	c_2^{asset}		2.93	0.71
			(0.61)			(0.35)	(0.45)
$c_{1,UPA}^0$			1.80	c_3^{asset}		3.20	1.13
			(0.49)			(0.37)	(0.46)
$c_{2,UPA}^0$			3.57	c_4^{asset}		2.27	0.38
			(0.54)			(0.33)	(0.40)
$c_{3,UPA}^0$			3.19	c_1^{Muslim}		0.10	0.06
			(0.54)			(0.44)	(0.44)
$c_{4,UPA}^0$			2.84	c_2^{Muslim}		-0.17	0.13
			(0.50)			(0.52)	(0.60)
c_1^{educ}		1.71	0.05	c_3^{Muslim}		-1.51	-1.28
		(0.33)	(0.48)			(0.46)	(0.48)
c_2^{educ}		2.68	0.59	c_4^{Muslim}		-0.31	-0.08
		(0.45)	(0.58)			(0.44)	(0.46)
c_3^{educ}		2.80	0.32				
		(0.44)	(0.58)				
Log likelihood					-1187.01	-859.85	-781.58

Notes: $b^{w,s}$ denotes the parameter of the interaction between w and s. Types 1-4 are the educated, uneducated, Muslim, and criminal types, respectively. Number of markets: 434.

Table A.14: Party objective function estimates: Constituency-level heterogeneity

	(1)	(2)	(3)	(4)		(1)	(2)	(3)	(4)
b^w	4.36	4.40	4.27	4.40	c_2^{educ}	0.39	0.69	0.68	0.55
	(0.84)	(0.85)	(0.83)	(0.86)		(0.54)	(0.55)	(0.55)	(0.56)
b^s	-11.31	-12.15	-11.43	-12.59	c_3^{educ}	0.26	0.43	-0.46	0.43
	(1.22)	(1.27)	(1.23)	(1.31)		(0.55)	(0.55)	(0.55)	(0.56)
c_1^*	1.10	0.62	5.81		c_4^{educ}	0.13	0.31	0.31	0.28
	(0.47)	(0.44)	(1.87)			(0.45)	(0.45)	(0.45)	(0.46)
c_2^*	1.99	0.01	4.18		c_1^{crime}	0.89	0.95	0.83	0.97
	(0.56)	(0.52)	(2.29)			(0.32)	(0.31)	(0.31)	(0.32)
c_3^*	0.99	0.85	3.34		c_2^{crime}	0.72	0.72	0.69	0.73
	(0.55)	(0.53)	(2.21)			(0.39)	(0.39)	(0.39)	(0.40)
c_4^*	1.50	1.58	5.17		c_3^{crime}	0.41	0.46	0.40	0.49
	(0.48)	(0.47)	(1.93)			(0.39)	(0.38)	(0.38)	(0.39)
$c_{1,NDA}^0$	3.16	3.30	-1.58	4.40	c_4^{crime}	-0.19	-0.08	-0.22	-0.04
	(0.55)	(0.56)	(1.59)	(1.35)		(0.32)	(0.32)	(0.32)	(0.33)
$c_{2,NDA}^0$	3.45	3.91	0.20	5.86	c_1^{asset}	-0.14	0.68	0.92	0.91
	(0.53)	(0.56)	(1.95)	(1.51)		(0.46)	(0.36)	(0.38)	(0.38)
$c_{3,NDA}^0$	5.07	5.07	2.30	6.29	c_2^{asset}	-0.57	0.62	0.84	0.94
	(0.60)	(0.61)	(1.91)	(1.59)		(0.53)	(0.43)	(0.45)	(0.45)
$c_{4,NDA}^0$	4.23	4.15	0.11	5.02	c_3^{asset}	0.37	1.18	1.23	1.34
	(0.56)	(0.56)	(1.64)	(1.44)		(0.54)	(0.44)	(0.46)	(0.46)
$c_{1,UPA}^0$	2.34	2.24	-2.42	3.32	c_4^{asset}	-0.79	0.32	0.43	0.56
	(0.46)	(0.48)	(1.57)	(1.37)		(0.47)	(0.37)	(0.38)	(0.39)
$c_{2,UPA}^0$	3.50	3.63	0.17	5.55	c_1^{Muslim}	0.19	0.32	0.34	0.35
	(0.52)	(0.56)	(1.96)	(1.56)		(0.44)	(0.44)	(0.43)	(0.45)
$c_{3,UPA}^0$	3.54	3.31	0.83	4.54	c_2^{Muslim}	0.16	0.22	0.27	0.31
	(0.52)	(0.54)	(1.90)	(1.60)		(0.59)	(0.59)	(0.59)	(0.61)
$c_{4,UPA}^0$	3.40	3.05	-0.77	3.90	c_3^{Muslim}	-1.14	-0.99	-0.99	-0.91
	(0.47)	(0.49)	(1.62)	(1.45)		(0.48)	(0.48)	(0.48)	(0.49)
c_1^{educ}	0.01	0.21	0.12	0.17	c_4^{Muslim}	-0.08	0.18	0.06	0.24
	(0.44)	(0.43)	(0.44)	(0.44)		(0.46)	(0.46)	(0.45)	(0.47)
Log likelihood						-786.17	-778.51	-789.12	-774.30

Notes: Estimates of party objective functions allowing for additional heterogeneity. Columns 1-3 interact candidate types with the following variables, respectively: 1: indicator for election year 2014; 2: indicator for reserved constituencies; 3: rural population. c_k^* is the parameter on the interaction for type k. In column 4, we include interactions with fixed effects for constituency clusters following Bonhomme et al. (2022) (the parameters on these are not shown). Types 1-4 are the educated, uneducated, Muslim, and criminal types, respectively. Number of markets: 434.

7.2 Constituency-level heterogeneity

In Table A.14 we study the robustness of our payoff function estimates to various types of observable and unobservable heterogeneity. In column 1 we consider the possibility that selection decisions are systematically different across election years 2009 and 2014. We augment the fixed costs to $c_k^0 + c_k^*$, where c_k^* is an additional direct cost in 2014. The estimates suggest that parties had a higher cost of replacing incumbents, and higher costs of selecting Type 2 and 4 candidates in 2014 than in 2009. However, there is little change in other estimates relative to our main estimates in the paper. In particular, the estimates of b^w and b^s are only marginally different, and so too are the rankings of the direct costs c_k^0.

In column 2, we allow costs to depend on constituency reservation status. The parameters c_k^* now denote an additional direct cost in reserved constituencies. The only estimate that is significantly different from 0 (at the 5% level) is c_4^*: there is a higher cost of selecting the criminal type in a reserved constituency. Again, our results are qualitatively unaffected by allowing for this type of heterogeneity.

In column 3 we allow for the possibility that selection decisions may depend on the rural/urban split in the constituency. The parameters c_k^* now denote an additional direct cost from a higher share of rural population. The estimates of b^w and b^s are again only marginally different from the estimates in the paper. The cost estimates indicate that the cost of replacing an incumbent is larger in rural constituencies.

Finally, we consider the possibility that constituency level *unobservables* may confound our estimates. Including a full set of constituency dummies in the model is not an option (interacting these with type choices would result in 651 additional parameters). Instead, we follow the approach of Bonhomme et al. (2022), who propose to reduce the dimensionality of the fixed effects by clustering units based on observable characteristics into groups in a first step. The idea is that constituencies in the same cluster likely have similar *unobservables*. We use kmeans as proposed by Bonhomme et al. (2022) to cluster constituencies using the full set of constituency level observables in the data. The algorithm identifies four constituency types, and we include dummies for these types (interacted with candidate types) in column 4 of Table A.14. Controlling for constituency level unobserved heterogeneity yields similar b^w and b^s estimates to our main specification, and causes little qualitative change in the cost parameters.

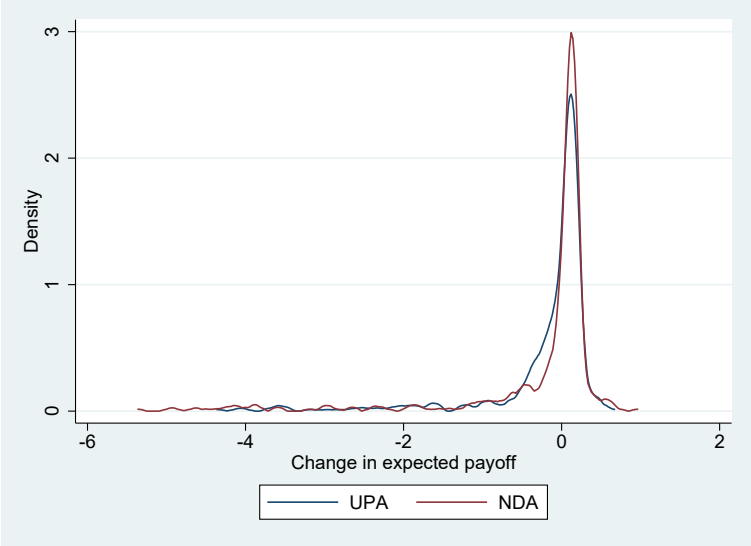

Figure A.4: Distribution of the change in parties' payoffs with a criminal ban

References

Andrews, Isaiah, Matthew Gentzkow, and Jesse M. Shapiro (2017): "Measuring the Sensitivity of Parameter Estimates to Estimation Moments," *The Quarterly Journal of Economics*, 132 (4), 1553–1592. 8, 11

Asher, Sam, Tobias Lunt, Ryu Matsuura, and Paul Novosad (2021): "Development research at high geographic resolution: an analysis of night-lights, firms, and poverty in India using the SHRUG open data platform," *The World Bank Economic Review*, 35 (4), 845–871. 3

Bajari, Patrick, Han Hong, and Denis Nekipelov (2013): "Game theory and econometrics: A survey of some recent research," . 13

Bonhomme, Stéphane, Thibaut Lamadon, and Elena Manresa (2022): "Discretizing unobserved heterogeneity," *Econometrica*, 90 (2), 625–643. 23, 24

Chhibber, Pradeep K and Rahul Verma (2018): *Ideology and identity: The changing party systems of India*, Oxford University Press. 7

Gandhi, Amit and Jean-François Houde (2019): "Measuring substitution patterns in differentiated-products industries," *NBER Working paper w26375*. 7, 9

Kitschelt, Herbert and Daniel M. Smidt (2013): "Democratic Accountability and Linkages Project (DALP)," https://sites.duke.edu/democracylinkage/. 7

Olea, José Luis Montiel and Carolin Pflueger (2013): "A Robust Test for Weak Instruments," *Journal of Business & Economic Statistics*, 31 (3), 358–369. 8

Rousseeuw, P.J. (1987): "Silhouettes: a Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Computational and Applied Mathematics*, 20, 53–65. 16