Discrimination or Ambition: What Drives Underrepresentation in the U.S. House?

Anubhav Jha*

Click here for most recent version

Abstract

The underrepresentation of minorities in key government bodies persists across all democratic institutions. For the U.S. House, scholars have identified two leading causes: voter discrimination and election aversion (lower political ambition), which are difficult to isolate from one another. To address this, the paper structurally estimates a model of political entry, voter discrimination, and campaign spending to separate the role of voter discrimination from that of election aversion in explaining underrepresentation. The framework also differentiates discrimination in primaries from that in general elections by modeling both stages. The model identifies election aversion by comparing general election outcomes (campaign spending and voting) between districts with only majority race (or gender) candidates and those with only minority candidates. When candidates on the ballot share the same identity, voters cannot discriminate based on identity. General election voter discrimination is then identified by comparing equilibrium outcomes between same-identity and mixed-identity districts. The dynamic structure of primaries followed by the general election allows me to account for candidate incentives and general election voter discrimination while recovering primary voter discrimination. I find that primary voter discrimination is the main driver of underrepresentation in the U.S. House. Although underrepresented groups show lower political ambition and face general election voter bias, these factors contribute minimally to underrepresentation. Policy counterfactuals show that a $150,000 campaign support subsidy during the primaries for underrepresented groups increases representation by 30% for Democrats and 177% for Republicans, while the same support in general elections has a negligible impact.

*Department of Politics, Princeton University. Email: aj7954@princeton.edu

I would like to thank Siwan Anderson, Thomas Fujiwara, Matias Iaryczower, Gleason Judd, Laura Karpuska, Kristopher Ramsey, Paul Schrimpf, and Francesco Trebbi for their comments and suggestions. I would also like to thank seminar and conference participants at International Conference on Game Theory (Stony Brook), Princeton Political Economy Seminar, and Structural Reading Group at Indian Statistical Institute-Delhi. Vaishnavi Singh provided excellent research assistance. All errors are my own.

This paper previously circulated under the name *"Screening for Influence: Latent Effects of Campaigning and Institutional Design in U.S. Congressional Races (2002-2018)"*

JEL Codes: D72, J15, J16, P0
Keywords: Elections, Voting Behavior, Economics of Gender and Race, Discrimination

1 Introduction

Minority groups have been consistently underrepresented in the U.S. legislature. Even today, with demographic shifts (Allen and Farley, 1986; Frey, 2018; Johnson and Lichter, 2016) and higher enfranchisement of voters (Cascio and Washington, 2014; Bernini et al., 2023; Schuit and Rogowski, 2017), a substantial disparity remains between the share of minorities in the general population and their representation in the legislature.[1] This representation gap extends beyond national and state-level institutions and exists deep within local governing bodies, including cities where minority populations constitute more than 50% of the local demographic (Trebbi et al., 2008; Ricca and Trebbi, 2022). Underrepresentation is more severe for female politicians than for racial minorities (Lawless and Pearson, 2008; Lawless and Fox, 2010; Burrell, 2014), with the average proportion of female House representatives being 16.6%. These gaps also exist for primary winners running for the U.S. House elections.[2]

Scholars in Political Economy, Political Science, and other related fields have often considered voter discrimination and election aversion as the leading causes of underrepresentation (Anzia and Berry, 2011; Ashworth et al., 2024; Fox and Lawless, 2004, 2011; Lawless and Fox, 2010; Kanthak and Woon, 2015). Separating the effects of these two mechanisms on the political selection of underrepresented groups is difficult. Anzia and Berry (2011) point out that voter discrimination can also lead to election aversion due to lower anticipatory gains from running for office. Ashworth et al. (2024) show that studies using close election Regression Discontinuity Design will not succeed in separating the effects of these mechanisms from one another.[3] Although these insights are developed to better understand the Political Economy of Gender, they also apply to the Political Economy of race.

To address the aforementioned challenges, this paper presents a comprehensive model of political entry, voter discrimination, and campaign spending that allows for the decomposition of the effects of voter discrimination, interest group discrimination, party leader-

[1] For example, while African Americans constituted 11.3% of the American population from 2002 to 2022, they held only 4% of the seats in the U.S. House.

[2] For instance, among the primary winners, 5.14% were African Americans, 4% were Hispanics, 2.7% belonged to other minority races, and 18% were female. This suggests that underrepresentation occurs not only in the U.S. House but may begin at the primary stage itself.

[3] Authors show that under differential costs/election aversion, conditional on close elections women will perform better than men in office. Additionally, they also show that under voter discrimination the same result can hold. Therefore, both mechanisms lead to women possessing higher quality than men conditional on close elections.

ship discrimination, and differential political ambitions on political selection.[4] The paper also distinguishes the discrimination faced by candidates in the primaries from that in the general elections by modeling both stages separately while dynamically linking them to one another. The model identifies election aversion by exploiting the differences in equilibrium outcomes across congressional districts where the competing candidates belong to the majority race (or gender) and those where the candidates belong to the same minority race (or gender) respectively.[5] When the candidates on the ballot share the same social identity, voters cannot vote based on identity, and therefore, the channel of voter discrimination is absent. Voter discrimination is recovered by comparing equilibrium outcomes between districts where the competing candidates share the same race (or gender) to those where they do not.[6] Finally, the nested structure of primaries followed by the general election allows me to account for candidate incentives in the primary and recover the degree of voter discrimination in the primary stage.

The proposed model consists of two main stages: the Entry and Primary (EP) stage and the General Election (GE) stage. In the EP stage, I assume that, in each party-specific primary, a continuum of candidates with a given mass simultaneously decides whether to contest in the primaries. Primary winners, selected through a large Tullock contest success function, then proceed to the GE stage. The GE stage is modeled as a Tullock contest game, where two alliances (one representing each primary winner) compete for the congressional district seat. Each alliance comprises of three individual players: the primary winner, a representative interest group, and the party leadership. Each player simultaneously decides on their level of spending to influence the contest function in their favor. Once these spending decisions are made, the Tullock contest success function determines the winner.

Potential candidates in the EP stage make costly entry decisions. These decisions are influenced by multiple forces, which can be categorized into two elements: the payoff they receive from winning the primary and the likelihood of winning the primary. This also introduces a challenge for identification, as shifting these elements in opposite directions can yield identical outcomes. For example, the number of primaries won by a social group with a lower preference for office but that is more favored by primary voters might be the same as if that group had a higher preference for office but was less favored by primary voters. This

[4] An overlooked potential cause of underrepresentation is discrimination by party leadership and interest groups, which actively participate in political campaigning in various ways (Bombardini and Trebbi, 2011; Cox, 2022). These groups can aid specific candidates and improve their chances of winning, while ignoring other candidates. If there is discrimination from interest groups and party leadership, some potential candidates from underrepresented groups may never choose to enter politics due to the anticipation of an expensive campaign. This mechanism is also hard to separate from election aversion.

[5] Interest group discrimination and party leadership discrimination are also recovered using these differences as I observe campaign spending decisions by these entities as well.

[6] As election aversion is known, therefore the differences must arise from voter discrimination.

poses a challenge for differentiating the mechanism of election aversion from that of voter discrimination if one only observes the social identity of primary winners. This problem can be solved if one knows the payoffs candidates receive from winning the primaries; then primary voter discrimination is the only unknown factor that influences the equilibrium outcomes in the EP stage. Changes in primary voter discrimination will lead to shifts in the share of primary winners from underrepresented groups and, therefore, will be recoverable. However, if the payoffs for primary winners are unknown, then the share of winners from underrepresented groups can only help identify the overall relationship between the payoffs received from winning the primary and the likelihood of winning, without being able to isolate these two factors.

The estimation of the GE stage recovers the payoffs candidates receive from winning the primary, but general election voter discrimination must still be identified separately from election aversion. To clarify how this is done, consider a simpler case where interest groups and party leadership do not campaign. Further, consider two congressional districts, A and B. In district A, two white candidates compete, and in district B, two black candidates compete. Note that voter discrimination is not a factor in determining the two-party vote shares of the candidates in these two districts, as in district B, the discrimination cancels out. However, the difference in the amount of effort (spending in our context) that candidates exert across these two districts depends on the degree of election aversion. Now, consider district C, where a black and a white candidate compete. Given that we know the election aversion, the differences in vote share and campaign spending between A and C (and also between B and C) will determine the degree of voter discrimination. Adding interest groups and party leadership introduces further complications, as there are now strategic complementarities whose effects need to be accounted for. This is where the structural model becomes useful, as it explicitly models these features and recovers the underlying election aversion, voter discrimination, and discrimination by interest groups and party leadership. This argument is demonstrated formally in Appendix B, while incorporating interest group, party leadership, and unobserved valence of candidates.[7]

For the empirical exercise, I use data from the Federal Election Commission (FEC) to obtain general and primary election outcomes. Additionally, I use independent expenditure data from the FEC, along with data provided by the Wesleyan Media Project and Wisconsin Advertising Project, to recover the spending decisions made by candidates, interest groups, and party leadership. Census and ACS data are used to recover congressional district demo-

[7]Another challenge in the GE stage is the issue of selection. Candidates who win primaries possess higher valence (unobserved) than those who lose, and failing to account for this selection would result in unreliable estimates. I identify candidate valence by exploiting comovements in spending decisions of candidates, interest groups, and party leadership. This is similar to the identification of the "common good" by using comovements in voting decisions for California ballot propositions in Matsusaka and Kendall (2021).

graphics. Finally, to recover the race of candidates, I use prompts to GPT-4, where I provide the full names of candidates, the year of the election contested, and the congressional district (with the state), and ask it to classify the candidate into their race/ethnicity category.[8] The gender and platform positions of the candidates are obtained from Bonica (2019).

Estimating the GE stage of the model reveals that interest groups prefer African American candidates over candidates from non-Black racial groups. Moreover, interest groups also show a stronger preference for female candidates over male candidates by −3.56 ln USD (0.335).[9] Party leadership favors African American candidates over White candidates and those from other races, with no significant gender bias. In terms of candidate preferences, Hispanic and White Americans value winning office more than African American candidates by 2.67 ln USD (0.77) and 1.91 ln USD (0.79), respectively. Male candidates demonstrate a higher preference for office than female candidates by 1.44 ln USD (0.18). This confirms the presence of election aversion among women candidates, as studied in the literature. General election voters prefer White and Hispanic candidates over African American and other race candidates. Additionally, male candidates are favored over female candidates by general election voters, confirming the arguments posed by Anzia and Berry (2011).

The estimation of the EP stage shows that primary voters have preferences for certain racial groups over others. Starting with Republican primary voters, I find that they prefer White candidates and candidates from other races more than Hispanic and African American candidates.[10] For Democratic primary voters, I find that White candidates are most preferred, followed by Black candidates, with Hispanic candidates and candidates from other races ranked lowest.[11] Primary voters from both parties show a preference for male candidates over female candidates, with estimates indicating a stronger preference among Republicans than among Democrats.

Although discrimination occurs at both stages of the election process, and candidates differ in political ambition, these factors do not always lead to significant underrepresentation, as equilibrium responses can either offset or amplify their effects. To analyze this, I make each player (or contest function) indifferent to candidate race and gender one at a time, then calculate equilibrium outcomes across all districts. I then compare the resulting race and gender shares at each stage with the observed shares to assess the impact of discrimination (or election aversion) by each entity on underrepresentation. For example,

[8] In Appendix C, I assess the quality of predictions made by GPT-4 for general election winners with CQPress' candidate biographical information. 90% of GPT-4's predictions match with CQPress.

[9] Candidate, interest group, and party leadership preference parameters are measured in units of natural log (ln) of USD.

[10] The estimated preference for Republican primary voters can be summarized as White $\sim_{Rep}$ Other race $\succ_{Rep}$ African American $\sim_{Rep}$ Hispanic.

[11] This preference can be summarized as White $\succ_{Dem}$ Black $\succ_{Dem}$ Hispanic $\sim_{Dem}$ Other race.

to isolate the effect of primary voter discrimination on underrepresentation, I modify their parameters to ensure primary voter preferences are constant across race and gender while keeping discrimination by other entities unchanged. This allows me to evaluate political selection at both stages of the election if primary voter discrimination is absent. Similarly, to examine patterns of political selection where general election voters do not discriminate, I adjust their preference parameters while keeping those of other entities constant.

My findings show that primary voter discrimination accounts for at least 83% and 70% of racial minority underrepresentation among Democrats and Republicans, respectively. Furthermore, primary voter discrimination explains over 90% of female underrepresentation for both parties in the U.S. House. In contrast, general election voter discrimination accounts for only 0.6% and 6% of racial minority underrepresentation for Republicans and Democrats, respectively. For female underrepresentation, these figures are 0.45% and 5%, respectively. Finally, election aversion contributes only marginally to underrepresentation in the U.S. House. These findings underscore that primary voter discrimination is the main driver of political underrepresentation in the U.S. House.

The paper also analyzes three policies aimed at increasing minority representation in the U.S. House. The first and second policy experiments involve providing campaign subsidies at the general election stage and the primary stage, respectively.[12] If campaign subsidies are provided at the general election stage, they result in negligible changes in representation. This is unsurprising, as discrimination at the general election stage contributes minimally to underrepresentation. However, subsidies provided at the primary stage lead to substantial improvements in representation. For example, a $150,000$ USD campaign subsidy for candidates from underrepresented groups results in a 177% improvement for Republicans and a 30% improvement for Democrats.

The third policy counterfactual is a reservation/quota policy that reserves a proportion of seats for underrepresented race-gender pairs (Desai et al., 2024; Clayton, 2021; Rosen, 2017). Such a policy is, however, extremely unlikely to pass in the United States, as it is unconstitutional. Nevertheless, many countries implement this policy, making it valuable to study in contexts where it is not yet adopted. This approach allows us to examine the political selection implications of quota/reservation policies, which may be harder to analyze in countries already employing such policies. I find that achieving a 68% share of House seats, matching the average share of the non-male-white population in my sample, would require reserving 50% to 60% of seats exclusively for underrepresented groups to contest. A 20% quota can significantly improve representation—Democrats by 37% and Republicans by 101%. These gains come with a slight increase in polarization (1-2%) and modest improvements in can-

[12] U.S. presidential candidates who meet certain conditions are eligible for campaign funding from the government. For details on public funding of presidential elections, see FEC webpage.

didate quality.[13]

This paper contributes to several strands of literature. First, it adds to the body of work modeling and estimating elections and campaigning (Cox, 2022; Kawai and Sunada, 2022; Iaryczower et al., 2022; Diermeier et al., 2005; Iaryczower et al., 2022; Strömberg, 2008; Bombardini and Trebbi, 2011; Acharya et al., 2022; Huang and He, 2021; Jha, 2023). The works most closely related to the model presented here are Kawai and Sunada (2022) and Cox (2022). Kawai and Sunada (2022) offers a dynamic model that incorporates candidates' entry and spending decisions. Additionally, it models both inter- and intra-election dynamics. Cox (2022) models candidate entry and platform decisions, as well as primary and general elections, along with PAC/Super-PAC spending decisions. This paper expands on these works by modeling and structurally estimating political selection based on the race and gender of candidates. Additionally, this paper recovers the preferences of candidates, interest groups, party leadership, and voters over the social characteristics of candidates, an important component missing in the literature on U.S. electoral campaigns.[14]

The paper also contributes to the literature on political selection (Acemoglu et al., 2010; Avis et al., 2022; Besley, 2005; Besley et al., 2010; Dal Bó et al., 2017; Dal Bó and Finan, 2018; Hirano et al., 2014). Given the vastness of the literature, I focus on two closely related works.[15] Dal Bó et al. (2017) documents patterns of political selection across various socio-economic dimensions and provides evidence of an "inclusive meritocracy" in Sweden. Hirano et al. (2014) highlights the critical role primaries play in selecting good versus bad politicians, depending on whether they occur in swing or stronghold congressional districts. This paper contributes to the literature by isolating and quantitatively analyzing the extent to which voter preferences, interest group preferences, party leadership preferences, and candidates' political ambitions contribute to creating disparities between political elites and the general population.

The paper also contributes to the literature on the Political Economy of Gender (Anzia and Berry, 2011; Ashworth et al., 2024; Fox and Lawless, 2004, 2011; Lawless and Fox, 2010; Kanthak and Woon, 2015) and Race (Trebbi et al., 2008; Ricca and Trebbi, 2022; Trounstine and Valdini, 2008; Trounstine, 2010; Shah et al., 2013; Shah, 2014; Marschall et al., 2010; Davidson and Korbel, 1981; Warshaw, 2019) in the United States. Ashworth et al. (2024) pro-

[13] Desai et al. (2024) estimate a model of discrimination and electoral accountability with probabilistic term limits to directly assess a polity where quotas exist. They also find that quotas are essential in maintaining representation of women candidates. There are also welfare losses, largely due to taste based discrimination, but these can be improved by mitigating perverse effects of term limits.

[14] Iaryczower et al. (2022) study Brazilian elections and recover voter preferences over the valence characteristics of candidates, providing a supply-side perspective on platform choice and estimates of trade-offs that vary by the valence characteristics of candidates.

[15] See Dal Bó and Finan (2018) and Besley (2005) for detailed reviews.

vides a model of political entry that accounts for gendered differences in election aversion and voter discrimination. This paper extends that model by incorporating discrimination by interest groups and party leadership, and further distinguishes between primary voter discrimination and general election voter discrimination. The literature on the Political Economy of Race has focused on representation in local governments, exploiting rich variation in demographics, city council composition, and electoral rules. Overall, this paper contributes to both strands of literature by analyzing national-level politics and revealing that primary voter discrimination is the leading cause of underrepresentation for both women and racial minority candidates.[16]

2 Model

The game takes place in two stages. In the first stage, I model the political entry decisions and primary races. I assume that primaries occur sequentially in a random order. In each primary, a continuum of agents simultaneously decides whether to enter the race. The winner is determined by a large Tullock contest success function. In the second stage, the general elections take place, where the primary winners, representative interest groups, and party leadership from both parties make their spending decisions. Following these decisions, the election winner is determined according to a Tullock contest success function.

Potential candidates are forward-looking; they internalize their future chances of winning, campaign costs, and the support they may receive from party leadership and interest groups. Moreover, these candidates are subject to both inter-party and intra-party competition. The model I propose captures the dependence of political selection not only on the preferences of voters and the candidates themselves, but also on those of interest groups and party leadership. Furthermore, it allows one to isolate the influence of primary voters on political selection from that of general election voters.

2.1 Second Stage: General Elections

I follow Kang (2016) to model the general election as a contest between two candidates. Each candidate (C) can choose to spend dollars. In addition to the candidate we have four

[16] Lawless and Pearson (2008) documents negligible differences in the winning chances of women candidates, conditional on those women who choose to enter. This sub-sampling introduces selection bias, as women candidates who choose to contest elections are likely to possess higher unobserved valence compared to women who do not contest and the average men who do contest (Anzia and Berry, 2011; Ashworth et al., 2024). The model I propose addresses this selection bias by modeling the entry decision of candidates. In equilibrium, candidates from discriminated groups, conditional on contesting, either possess higher valence, platform positions closer to primary median voters, or a combination of both.

additional players in this stage: the representative interest group (IG) aligned with each candidate's party and each candidate's party leadership (PL). These four players may also spend dollars in support of their respective candidates. The electoral outcome is modeled as a Tullock contest success function. Let $d \in \{1, 2, \ldots, D\}$ be an arbitrary district, $i \in \{1, 2\}$ be an arbitrary candidate, and let $P_{i,d} \in \{R, D\}$ denote the party of the candidate. Then probability that i wins the general election is given by:

$$Prob\left[i \text{ wins} \middle| \left\{\mathbf{s}_{i,d}, Q_{i,d}, p_{i,d}, \xi_{q,i,d}\right\}_{i\in\{1,2\}}; X_d\right] = \frac{h + \sum_{l\in\{C,IG,P\}} \beta \cdot s_{i,d,l}^{\gamma}}{1 + \sum_{j\in\{R,D\}} \sum_{l\in\{C,IG,P\}} \beta \cdot s_{j,d,l}^{\gamma}}, \tag{2.1}$$

where $s_{i,d,l}$ denotes spending made by $l \in \{C, IG, PL\}$ aligned with candidate i in district d, $Q_{i,d}$ denotes the valence characteristics of candidate i in district d, $p_{i,d}$ is candidate's platform position, β is effectiveness of spending, and γ is parameter that captures decreasing returns to political spending. The term h is a function of median voter ideology $\equiv X_d'\beta_{I,v}$ (X_d is a vector of congression district characteristics), vector of candidate characteristics $Q_{i,d}$, platform positions $p_{i,d}$, and unobserved valence term $\xi_{q,i,d}$ for each $i \in \{1, 2\}$. I parameterize the function as followed:

$$h(\mathbf{Q}_d, \mathbf{p}_d, X_d, \boldsymbol{\xi}_q) = \frac{\exp\left(Q_{i,d}' \cdot \beta_{q,v} - w_{I,v} \cdot \left(p_i - X_d' \cdot \beta_{I,v}\right)^2 + \xi_{q,i,d}\right)}{\sum_{j\in\{R,D\}} \exp\left(Q_{j,d}' \cdot \beta_{q,v} - w_{I,v} \cdot \left(p_j - X_d' \cdot \beta_{I,v}\right)^2 + \xi_{q,j,d}\right)}, \tag{2.2}$$

If candidate i wins the election, then players $l \in \{C, IG, PL\}$ aligned with i receive the following payoff

$$\log U_{i,d,l} = Q_{i,d}'\beta_{q,l} - w_{I,l} \cdot (p_i - I_{i,d,l})^2 + \xi_{q,i,d} + \xi_{cost,i,d,l} \tag{2.3}$$

where $\beta_{q,l}$ is an $l-$specific vector of coefficients associated with characteristics $Q_{i,d}$. The term $I_{i,l,d}$ is preferred platform position of $l \in \{C, IG, PL\}$ aligned with i in district d. For $l = C$, I assume that $I_{i,d,C} = p_{i,d}$ that $p_{i,d}$ is the true platform position of the candidate. For $l = IG$, I assume that $I_{i,d,IG} = I_{IG}$ that is the representative interest group have the same preferred ideology positions. Note that IGs are not always party loyal therefore allowing for party specific ideology would lead to superficially polarized representative interest groups which is not the case. For $l = PL$, $I_{i,d,PL} = I_{R,PL} \times \mathbb{1}\{P(i,d) = R\} + I_{D,PL} \times \mathbb{1}\{P(i,d) = D\}$. The term $\xi_{q,i,d}$ denotes candidate i's unobserved valence term and $\xi_{cost,i,d,l}$ denotes unobserved idiosyncratic cost shock that varies across i and l.

For each district d, all players simultaneously decide whether to campaign for candidate i or not. This decision is denoted by $a_{i,d,l}$ for each l aligned with i in district d. These decisions are then observed by all players. Then the participating players, for whom $a_{i,d,l} = 1$ simultaneously make the intensive margin decision, where they choose the amount of money, $s_{i,d,l}$,

to spend in support of candidate i. The spending problem is defined as,

$$\max_{s_{i,d,l}\in(0,\infty)} U_{i,d,l} \cdot \frac{h + \sum_{k\in\{C,IG,P\}} \beta \cdot s_{i,d,k}^{\gamma}}{1 + \sum_{j\in\{R,D\}} \sum_{g\in\{C,IG,P\}} \beta \cdot s_{j,d,g}^{\gamma}} - s_{i,d,k}, \tag{2.4}$$

where the marginal cost of spending is assumed to be 1 since it is not separately identified from preferences $U_{i,d,l}$. Now I define the program that players solve to decide whether they should participate in the campaign or not. Let $\mathbf{W}_d = \left(\{Q_{i,d}, p_{i,d}\}_{i\in\{1,2\}}, X_d\right)$ is the vector of candidate and congressional district characteristics. Let the term $a_{i,d,k}$ denote the campaign participation decision made by player k aligned with i in district d. Then, the participation problem can be defined as

$$V_l^{II}(\mathbf{W}_d, \boldsymbol{\xi}_d) = \max_{a_{idl}\in\{0,1\}} \hat{U}_{i,d,l}\left(a_{i,d,l}; \mathbf{a}_{i,d,-l}, \mathbf{a}_{-i,d}, \mathbf{W}_d, \boldsymbol{\xi}_d\right), \tag{2.5}$$

where $\hat{U}_{i,d,l}\left(a_{i,d,l}; \mathbf{a}_{i,d,-l}, \mathbf{a}_{-i,d}, \mathbf{W}_d, \boldsymbol{\xi}_d\right)$ is the equilibrium payoff players get in the spending stage.

2.2 First Stage: Entry and Primaries

The primaries take place sequentially, where nature picks D as the first mover party with probability half. Suppose, R is the second mover party and characteristics of the primary winner for D are $E_{j,d}^{O} = \left(Q_{j,d}, p_{j,d}, \xi_{q,j,d}\right)$. There is a continuum of potential R candidates of mass λ who simultaneously decide whether they should contest in the primaries or not. This continuum of potential candidates are distributed according to density $f_R(Q,p,\xi) = f_q(Q)\cdot f_{R,I}(p)\cdot f_{R,\xi}(\xi)$. Let $\pi^s\left(Q,p,\xi; E_{j,d}^{O}\right)$ be the probability with which an R candidate with characteristics Q,p,ξ chooses to enter when the winner of D is observed to have characteristics $Q_{j,d}, p_{j,d}, \xi_{j,d}$. Then I define the density with which a candidate with characteristics Q_i, p_i, ξ_i wins the primary as:

$$\Pi_{s,R}^{W}\left(Q_i, p_i, \xi_i; \pi^s, E_{j,d}^{O}; X_{d,Prim}\right) = \frac{e^{Q_i'\delta_q - \left(p_i - X_{d,Prim}'\delta_I\right)^2 + \xi_i}}{\int_{Q,p,\xi} e^{Q'\delta_q - \left(p - X_{d,Prim}'\delta_I\right)^2 + \xi} \cdot \pi^s\left(Q,p,\xi,E_{j,d}^{O}\right) dF_R(Q,p,\xi)}, \tag{2.6}$$

where δ_q is a vector of coefficients associated with candidate characteristics, $X_{d,Prim}$ is a vector of congressional district characteristics which also includes controls for the type of primary race, and δ_I is the corresponding vector of coefficients. The numerator of the expression represents the contribution a potential candidate's characteristics make to their probability of winning. The denominator represents the contribution made by characteristics of everyone else who chooses to contest the primary. The problem solved by a potential candidate is then defined as:

$$\max_{\pi\in\{0,1\}} \left(\mathbb{E}\left[V_C^{II}(\mathbf{W}_d, \boldsymbol{\xi}_d) \middle| Q_i, p_i, \xi_i, E_{j,d}^{O}, X_d\right] \cdot \Pi_{s,R}^{W}\left(Q_i, p_i, \xi_i; \pi^s, E_{j,d}^{O}; X_{d,Prim}\right) - \kappa\right) \cdot \pi, \tag{2.7}$$

where κ is the entry cost that candidate face while making the entry decision. The variable π denotes an arbitrary entry decision. Note that $\pi^s\left(Q,p,\xi;E^O_{j,d}\right) = \pi^*$ where, π^* solves the problem 2.7 for (Q,p,ξ) when the winner of the first mover primary is E^O_{jd}. In this problem, the candidate is maximizing the product of the expected GE payoff, $\mathbb{E}\left[V^{II}_C(\ldots)|\ldots\right]$, and the candidate's primary race win density.

The mathematical problem that potential candidates for the first mover primary solve is similarly defined. Let $\pi^f(Q,p,\xi)$ be a potential candidate's equilibrium entry decision. Let $\Pi^W_{f,D}\left(Q_i,p_i,\xi_i;\pi^f,X_{d,Prim}\right)$ be their density of winning the primary, which is defined below

$$\Pi^W_{f,D}\left(Q_i,p_i,\xi_i;\pi^f;X_{d,Prim}\right) = \frac{e^{Q_i'\delta_q-\left(p_i-X'_{d,Prim}\delta_I\right)^2+\xi_i}}{\int_{Q,p,\xi} e^{Q'\delta_q-\left(p-X'_{d,Prim}\delta_I\right)^2+\xi}\cdot\pi^f(Q,p,\xi)\,dF_D(Q,p,\xi)}, \tag{2.8}$$

where δ_q, $X'_{d,Prim}$, and δ_I have the same definition as before. The potential candidates solve the following problem:

$$\max_{\pi\in\{0,1\}} \left(\mathbb{E}\left[V^{II}_C(\mathbf{W}_d,\boldsymbol{\xi}_d)\middle|E^O_i=(Q_i,p_i,\xi_i),X_d,X_{d,Prim}\right]\cdot\Pi^W_{f,D}(Q_i,p_i,\xi_i;\pi^e;X_{d,Prim})-\kappa\right)\cdot\pi, \tag{2.9}$$

where the expectations $\mathbb{E}\left[V^{II}_C(\mathbf{W}_d,\boldsymbol{\xi}_d)\middle|E^O_i=(Q_i,p_i,\xi_i),X_d,X_{d,Prim}\right]$ captures the fact that potential candidates anticipate the competition they will face in the general election from the other primary's winner.

2.3 Equilibrium

I assume in Assumption 2.1 that unobserved shocks are public knowledge for the players in the game. These shocks consist of two parts: the first are cost shocks, which are experienced by candidates, interest groups, and party leadership, capturing unobserved variations in spending costs faced by these players. The second part consists of unobserved valence shocks, which capture candidate characteristics such as charisma, oratory skills, public image, personality traits, and other qualities that the econometrician does not measure but that are crucial in explaining electoral outcomes. Moreover, these unobserved valence shocks affect the marginal benefits that candidates, interest groups, and party leadership derive from winning, not only by altering the probability of victory, but also by increasing or decreasing the value of the contested seat. Ultimately, this ensures that the spending decisions made by candidates, interest groups, and party leadership are implicit functions of unobserved valence. As a result, naive regression models that fail to account for this endogeneity cannot accurately capture the effect of spending on a candidate's win probability.

Assumption 2.1 *In the general election stage,* $\xi_d = \left(\{\xi_{q,i,d},\xi_{cost,i,d}\}_{i\in\{R,D\}}\right)$ *is public knowledge but unknown to the econometrician.*

The decision to participate in campaigning, made by candidates, interest groups, and party leadership, results in multiple equilibria. Therefore, without additional assumptions, the model does not yield a unique prediction for equilibrium outcomes (Bajari et al., 2010; Tamer, 2003). I assume that only pure strategy equilibria are played and that the probability of an equilibrium being selected is proportional to the sum of payoffs for all players. Consequently, Pareto superior equilibria—where the total payoff for all players is higher—are more likely to be selected than inferior ones, though Pareto inferior equilibria may still be selected with positive probability.[17] This is a departure from Kang (2016), where it is assumed that only the Pareto optimal equilibrium—the equilibrium where the sum of payoffs is maximum—is played.

Assumption 2.2 *In the general election stage, specifically the participation stage, pure strategy equilibrium is played. Moreover, in case of multiplicity, an equilibrium is randomly picked where the probability is proportional to sum of payoff of all players.*

I impose parametric assumptions on potential candidate distributions F_R and F_D. First, I assume $f_{p,\xi}$ is identical across the parties and follows a normal distribution with mean μ_ξ and standard deviation σ_ξ. The distribution of observed valence characteristics, $f_q(Q)$, is also assumed to be identical across parties. Assuming that the valence distribution is identical across parties allows for endogenous disparities in valence characteristics to emerge across parties. If we observe more representation for a particular social group in one party than the other, this difference will arise endogenously within the model and can be explained by demand-side factors, such as voter preferences (contest functions), IG/party leadership preferences, or it may be due to the fact that a social group does not derive sufficient value from winning under a given party's platform (explained by candidate preferences). I assume that $f_{P,p}$ is party-specific and equal to a normal probability density function with mean $\mu_{p,P}$ and standard deviation $\sigma_{p,P}$.

Assumption 2.3 *I impose the following parameteric assumptions on F_R and F_D:*

(a) $\xi_{i,P} \sim N(\mu_\xi, \sigma_\xi)$ *for* $P \in \{R, D\}$

(b) $p_{i,P} \sim N(\mu_{p,P}, \sigma_{p,P})$ *for* $P \in \{R, D\}$

(c) $Q = (Q_{cont}, Q_{disc})$, *where* $Q_{cont} \in \mathbb{R}^{K_{cont}}$ *and* $Q_{l,disc} \in \boldsymbol{Q}_{l,disc}$ *for* $l = 1, 2, \ldots, K_{disc}$.

(d) $Q_{cont} \sim N(\mu_{cont}, \Sigma_{cont})$.

(e) $Prob[Q_{l,disc} = x] = q_{x,l,disc}$ *such that* $x \in \boldsymbol{Q}_{l,disc}$ *and* $\sum_{x \in \boldsymbol{Q}_{l,disc}} q_{x,l,disc} = 1$

The equilibrium in the spending (the intensive margin of campaigning) stage is unique

[17] Due to computational constraints that arise when solving for Nash equilibria in N-player discrete games, I focus only on pure strategy equilibria.

and has been proven for this specification in Kang (2016). Proposition 2.1 states the proposition.

Proposition 2.1 *A unique equilibrium exists in the campaign spending stage given a campaign participation profile.*

For proof, see Proposition 1 in Kang (2016). The equilibrium entry decisions for the second mover primary are stated in Proposition 2.2. The complexity of these entry decisions is greatly reduced when considering a continuum of candidates, as demonstrated in Proposition 2.2. The problem is broken down into two parts: first, to describe the best responses of candidates, π^s, given their types as a function of the overall competitiveness of the election. This competitiveness is given by the variable A, which is essentially the denominator of the winning density function in equation 2.6. The second part is to find the equilibrium level of competitiveness by ensuring that the integral of the product of the win density and entry decisions with respect to F_P equals 1. That is $\int_{Q,p,\xi} \Pi^W_{s,P} \times \pi^s dF_P(Q,p,\xi) = 1$. This can be rewritten as equation 2.13. This reduces the problem from solving for the infinite-dimensional object π^s to solving an equation of one variable.

Proposition 2.2 *Suppose the winner of first mover primary has characteristics $E^O_{j,d}$, then*

$$\pi^s\left(Q_i,p_i,\xi_i;E^O_{j,d},X_{d,Prim}\right)=\begin{cases}1 & \frac{\exp\left\{Q_i'\delta_q-\left(p_i-X_{d,Prim}'\delta_I\right)^2+\xi_i\right\}}{A\left(E^O_{j,d},X_{d,Prim}\right)}\cdot\mathbb{E}\left[V_C^{II}\left(\mathbf{W}_d,\boldsymbol{\xi}_d\right)\middle|Q_i,p_i,\xi_i,E^O_{j,d},X_d\right]\geq\kappa\\ 0 & otherwise\end{cases}, \tag{2.10}$$

where $A\left(E^O_{j,d},X_{d,Prim}\right)$ uniquely solves the following equation

$$\int_{Q,p,\xi} e^{Q_i'\delta_q-\left(p_i-X_{d,Prim}'\delta_I\right)^2+\xi_i}\pi^s dF_P\left(Q,p,\xi\right)-A=0 \tag{2.11}$$

The proof first shows that the LHS of equation 2.13 is a continuous function of A. I then demonstrate that the LHS is monotonically strictly decreasing in A and as $A\to\infty$, the LHS approaches $-\infty$, while as $A\to 0$, the LHS approaches a positive number. Therefore, by the intermediate value theorem, there exists an $A\left(E^O_{j,d},X_{d,Prim}\right)=A^*$ that solves the equation. Similarly, the problem for the first mover is broken down and solved in Proposition 2.3.

Proposition 2.3 *The equilibrium entry decisions, π^f are given by*

$$\pi^f\left(Q_i,p_i,\xi_i;X_{d,Prim}\right)=\begin{cases}1 & \frac{\exp\left\{Q_i'\delta_q-\left(p_i-X_{d,Prim}'\delta_I\right)^2+\xi_i\right\}}{A\left(X_{d,Prim}\right)}\cdot\mathbb{E}\left[V_C^{II}\left(\mathbf{W}_d,\boldsymbol{\xi}_d\right)\middle|E_i^O=\left(Q_i,p_i,\xi_i\right),X_d,X_{d,Prim}\right]\geq\kappa\\ 0 & otherwise\end{cases}, \tag{2.12}$$

where $A\left(X_{d,Prim}\right)$ uniquely solves the following equation

$$\int_{Q,p,\xi} e^{Q_i'\delta_q-\left(p_i-X_{d,Prim}'\delta_I\right)^2+\xi_i}\pi^f dF_P\left(Q,p,\xi\right)-A=0 \tag{2.13}$$

Arguments similar to those used in proof A.2 also prove this proposition.

2.4 Illustration of Equilibrium Decisions: Toy Example

In this section, I discuss the selection patterns predicted by the Entry and Primary stage of the model in equilibrium. To illustrate this, consider the following simplifications: there is only one primary, and two social groups, Majority (Maj) and Minority (Min). Candidates from Maj receive a payoff of V_{Maj} if they win the primary, and Min candidates receive V_{Min} if they win. The weight that primary voters assign to Maj candidates is $\delta_{Maj} = 0$, while the weight for Min candidates is δ_{Min}. The entry costs for both candidate groups are identical, denoted by κ. The equilibrium entry decisions for Majority candidates are then given by:

$$\pi\left(Maj, p, \xi; A^*\right) = \mathbb{1}\left\{\exp\left\{-\left(p-I\right)^2 + \xi\right\} \cdot V_{Maj} > \kappa A^*\right\}, \tag{2.14}$$

where I is the median voter's preferred platform position, p is an arbitrary candidate platform position, ξ is an arbitrary unobserved candidate valence. For Min candidates the equilibrium entry decisions are given by:

$$\pi\left(Min, p, \xi; A^*\right) = \mathbb{1}\left\{\exp\left\{\delta_{Min} - \left(p-I\right)^2 + \xi\right\} \cdot V_{Min} > \kappa A^*\right\}. \tag{2.15}$$

From Proposition 2.2, the equilibrium level of competition (A^*) is given by solving the following equation

$$\begin{aligned} p_{Maj}\mathbb{E}_{\xi,p}\Big[\exp\left\{-\left(p-I\right)^2 + \xi\right\}\pi\left(Maj, p, \xi; A^*\right)\Big] + \\ (1-p_{Maj})\mathbb{E}_{\xi,p}\Big[\exp\left\{\delta_{Min} - \left(p-I\right)^2 + \xi\right\}\pi\left(Min, p, \xi; A^*\right)\Big] = A^* \end{aligned} \tag{2.16}$$

Now, consider majority candidates who are indifferent between contesting in elections and not contesting. These candidates are defined by the following equation.

$$\xi = \log\left(\frac{\kappa \cdot A^*}{V_{Maj}}\right) + (p-I)^2. \tag{2.17}$$

Similarly, the indifferent minority candidates are defined by the equation:

$$\xi = \log\left(\frac{\kappa \cdot A^*}{V_{Min}}\right) - \delta_{Min} + (p-I)^2. \tag{2.18}$$

The set of Majority candidates who choose to enter are given by, $\bar{E}^{Maj} = \{(p,\xi) : \xi \geq \log(\frac{\kappa \cdot A^*}{V_{Maj}}) + (p-I)^2\}$. Similarly, the set of minority candidates who choose to enter are given by $\bar{E}^{Min} = \{(p,\xi) : \xi \geq \log(\frac{\kappa \cdot A^*}{V_{Min}}) - \delta_{Min} + (p-I)^2\}$. Assume election aversion for minorities, $V_{Min} < V_{Maj}$, and taste based discrimination by voters, $\delta_{Min} < 0$. Then we get that $\bar{E}^{Min} \subset \bar{E}^{Maj}$, but the reverse does not hold. This implies that in equilibrium there is

a non-empty set of majority candidates, $\bar{E}^{Maj}/\bar{E}^{Min}$, who find it optimal to contest in the primaries but would choose not to contest if they were minority candidates. This set of candidates exhibit the behavior that equivalent to what political theory on election aversion predicts for candidates. However, as Anzia and Berry (2011) point out, this behavior will still hold in equilibrium even if I assume there is no election aversion, i.e. $V_{Min} = V_{Maj}$. Thus, voter discrimination and election aversion both contribute to hesitancy among minorities to participate in politics and can result in similar entry decisions. I illustrate these decisions in Figure 1, where the green-colored region represents candidates in the set $\bar{E}^{Maj}/\bar{E}^{Min}$ and the gray-colored region represents candidates in the set $\bar{E}^{Min}$.

3 Data

The dataset is constructed using multiple data sources. Electoral outcomes such as the number of entrants, primary winners, general election winners, and their vote shares were obtained from the Federal Election Commission's (FEC) website. Platform positions of candidates were obtained from Bonica (2019). To classify the candidates into their race-ethnic categories, I used GPT-4. I provided the full name of the candidate, the year, and the congressional district where the candidate ran for the U.S. House, then asked GPT-4 to classify the candidate into "non-Hispanic white," "non-Hispanic black," "Hispanic," and "Other."[18] I use FEC data on independent expenditures by non-candidate committees and total donations to candidate committees. This data is complemented by the Wesleyan Media Project and Wisconsin Advertising Project to construct expenditures undertaken by candidates, interest groups, and party leadership. Congressional district socio-economic outcomes were obtained from Census and ACS.

The General Election data covers the years 2002-2022, and the Primary Race data covers the years 2002-2020. The summary statistics are shown in Table 1. In the dataset, 87.7% of the 8,288 primary winners are White, followed by African Americans at 5.14%, Hispanics at 4.34%, and Other Races at 2.79%. Of these eight thousand primary winners, 18.8% are female candidates. The table also shows that the average spending by candidates in an election campaign was $944K$ USD, interest groups on average spent $156K$ USD, and party leadership spent an average of $100K$ USD on campaigning for candidates.[19] On average, there were 1.97 candidates entering a primary race, with variation ranging from some races having more than 10 candidates to races where a candidate proceeded to the General Elec-

[18] In Appendix C, I assess the quality of predictions made by GPT-4 for general election winners with CQPress' candidate biographical information. 90% of GPT-4's predictions match with CQPress. The confusion matrix is also provided. The worst prediction case is for Others. However, GPT-4 is not always incorrect for these mismatches, some entries at CQPress are incorrect as well which GPT-4 correctly predicted.

[19] These USD values are in 2022 dollars.

tion stage unopposed. There is also variation in the types of primary formats.

Minorities are underrepresented: The first empirical pattern I discuss is the underrepresentation of minorities in the U.S. House. This pattern is shown in Figure 4. On average, the proportion of African Americans, Hispanics, and Other races has been disproportionately lower than their share in the general population. This underrepresentation is not only present in Congress but also among Primary Winners, suggesting that the underrepresentation of these groups begins, at least in part, at the primary stage. Recently, there has been an improvement in the representation of African Americans and females, but only on the Democratic side. I document this pattern in Figure 5.

Demand Side Patterns of Underrepresentation of Minorities: Here, I analyze the demand for candidates using reduced form methods. It is important to recognize that this side consists not only of voters but also of interest groups and party leadership. To assess whether underrepresentation results from demand-side discrimination, it is necessary to separately recover the preferences of voters, interest groups, and party leadership. In this section, I will document certain patterns that reveal correlations between demand-side actions and candidates' characteristics. For this analysis, I run the following regression

$$Y_{c,t} = Q_c'\beta + X_{d(c),t}'\delta + \alpha_{s(c)} + \gamma_t + \epsilon_{ct}, \tag{3.1}$$

where c is a candidate, t is the year, $d(c)$ is the congressional district from which the candidate is contesting, $Y_{c,t}$ is an outcome variable of interest, $s(c)$ is the state from which the candidate is contesting, Q_c is a vector of candidate characteristics, $X_{d(c),t}$ is a vector of congressional district characteristics, $\alpha_{s(c)}$ is the state fixed effect, and γ_t is the year fixed effect. The above regression is estimated for spending by party leadership, spending by interest groups, and vote shares in the general elections.

Table 2 reports the results of the regression given in Equation 3.1. Consider columns (1) to (3), where I compare party spending on racial minorities with spending on white candidates, and I do not find any significant difference. The same pattern holds when considering spending by interest groups and vote shares. However, when the racial minority variable is broken down into individual races, such as African Americans, Hispanics, and other minority races (e.g., Asians, South Asians, Indigenous origin, etc.), distinct patterns emerge. I find that African American candidates are less likely to receive campaign support from party leadership and interest groups. The opposite holds for candidates of Hispanic origin. In terms of vote shares, I do not find any significant differences.

For female candidates, I find no evidence of discrimination by either party leadership or voters. The estimates for the female dummy are reported in columns (1) and (4) for party leadership, and in columns (3) and (6) for voters. For interest groups, however, I find that

they provide higher campaign support to female candidates than to male candidates, as reported in columns (2) and (5).

These empirical patterns show that the demand side is quite heterogeneous in terms of preferences over race and gender of candidates. While party leadership and interest groups show signs of stronger preferences for Hispanic candidates and weaker preferences for African American candidates, it is not clear whether these patterns are driven by correlations between candidate characteristics and competitive races or if they truly reflect the underlying preferences of party leadership and interest groups. Similarly, without accounting for the influence of spending decisions made by candidates, party leadership, and interest groups, it is difficult to determine whether the vote share regression estimates genuinely represent underlying preferences for candidate characteristics or if they are influenced by the correlation between spending decisions and candidate characteristics. To address this, I will leverage the structural model to uncover the true underlying preferences of these stakeholders regarding a candidate's race and gender attributes.

Candidate Characteristics, Party Affiliation, and Platform Positions: The combined effects of discrimination by voters, interest groups, and party leadership—along with their preferences over policy positions—and the differing political ambitions of minority candidates can lead to the selection of candidates whose platform positions diverge from those of majority candidates. To assess whether there are differences in the platform positions of candidates from underrepresented groups, I conduct the following regression analysis

$$Y_{c,t} = Q_c'\beta + X_{d(c),t}'\delta + \alpha_{s(c)} + \gamma_t + \epsilon_{ct}, \tag{3.2}$$

where $Y_{c,t}$ is an outcome of interest associated with candidate c in period t. The other variables remain the same as in Equation 3.1. Table 3 presents the results of this regression using a sample of primary winners.

As reported in column (1), racial minority candidates and female candidates are, on average, more liberal than their white and male counterparts, respectively. This finding remains consistent even when the racial minority category is disaggregated into African Americans, Hispanics, and other races, as shown in column (4). However, this pattern does not arise because racial minority and female primary winners are inherently more liberal, but rather because they are more likely to be affiliated with the Democratic Party than the Republican Party.

The within-party analysis provides additional insights. Racial minority candidates hold more liberal positions within the Republican Party and more conservative positions within the Democratic Party, while female candidates exhibit the opposite pattern, as indicated in columns (2), (3), (5), and (6). Therefore, within each party, there is no consistent evidence that underrepresented groups are more liberal on average. Furthermore, columns (7) and

(8) demonstrate that racial minority and female candidates are more likely to be affiliated with the Democratic Party.

This observation raises the question of why we see a higher number of general election minority candidates from the Democratic Party compared to the Republican Party. The parameter estimates from the structural model and the results of counterfactual experiments, discussed later, show that this phenomenon is driven by higher levels of discrimination based on race and gender among Republican primary voters compared to their Democratic counterparts. This higher degree of discrimination essentially leads to the sorting of underrepresented groups into the Democratic Party rather than the Republican Party.

4 Identification and Estimation

The game is estimated in two steps. The General Election (GE) stage is estimated first followed by the Entry and Primary (EP) stage.

4.1 General Election Stage

Identification: There are two important differences between Kang (2016) and the general election stage presented in this paper. The first difference is the presence of unobserved valence shocks, which introduce vertical differentiation among politicians. If such shocks were present in Kang (2016), they would be analogous to the vertical differentiation of policies for lobbyists. These unobserved valence shocks are identified through the comovement of spending decisions and general election voting outcomes. The second key difference is the incorporation of probabilistic equilibrium selection, allowing for the possibility that the data may be generated from Pareto-inferior equilibria.

To estimate the GE stage, I assume that valence shock, $\xi_{q,i,d}$, for primary winners in congressional district d are distributed according to the normal distribution with mean $X_d'\beta_{\xi_q}$ and standard deviation $\sigma_{\xi_q|PW}$.

Assumption 4.1 *Valence shocks for primary winners in congressional district d are distributed according to* $N\left(X_d'\beta_{\xi_q}, \sigma_{\xi_q|PW}\right)$.

I estimate the model using a penalized log-likelihood that accommodates equilibrium selection, a deviation from Kang (2016). I observe whether candidate $i = 1$ won in district d or not denoted by Y_d, whether l aligned with i participates in campaigning or not given by $A_{i,d,l}$. Moreover, I also observe spending levels by l aligned with i, denoted by $S_{i,d,l}$. Now given this, the set of parameters in GE stage is given by:

$$\Theta^{GE} = \left(\beta, \left\{\beta_{q,l}\right\}_{l\in\{v,C,IG,PL\}}, I_{ig}, \omega_{I,PL}, \beta_{I,v}, \beta_{\xi_q}, \sigma_{cost,C}, \sigma_{cost,IG}, \sigma_{cost,PL}, \sigma_{\xi_q|PW}\right), \tag{4.1}$$

where β is spending effectiveness, $\beta_{q,l}$ is coefficient associated with valence characteristics for l, I_{ig} is representative IG ideology, $\omega_{I,PL}$ is weight on ideology loss function for party leadership, $\beta_{I,v}$ is coefficient of X_d that explain variation median voter ideology, β_{ξ_q} are coefficients of X_d that explain variation in mean valence of primary winners in congressional district d, $\sigma_{cost,l}$ is standard deviation of cost shocks for player l, and lastly $\sigma_{\xi_q|PW}$ is the standard deviation of valence shocks of primary winners. Parameters $I_{D,PL}$ and $I_{R,PL}$ are calibrated to mean dime scores of congressional members as these are not separately identified from $\omega_{v,PL}$, and ω_v and ω_{IG} are calibrated to 1.0 as these are not separately identified from ideology estimates.

To see how the parameter $\beta_{\xi_q,j}$ and $\sigma_{\xi|PW}$ is identified note that preferences of player l aligned with i can be written as

$$log\left(U_{i,d,l}\right) = Q'_{i,d}\beta_{q,l} - w_{I,l}\cdot\left(p_i - I_{i,d,l}\right)^2 + \underbrace{X'_d\beta_\xi + \sigma_{\xi|PW}v_{i,d}}_{\xi_{q,i,d}} + \xi_{cost,i,d,l}, \tag{4.2}$$

where $v_{i,d}$ is a standard normal random variable. Note that if the number of players (l) were sufficiently large, one could reliably recover candidate valence using $i \times d$ fixed effects as one observes multiple players (candidate, interest groups, and party leadership) for each $i \times d$. However, due to the presence of only three players on each side and the corresponding sheer size of fixed effects that will be needed, I impose a distributional assumption on these shocks. Note that β_ξ and $\sigma_{\xi|PW}$ are common across all players, and therefore, the co-movements in spending decisions across players within congressional districts identify β_ξ and $\sigma_{\xi|PW}$.[20]

Identification of $\beta_{q,l}$ and $\beta_{q,v}$ follows the following argument. Consider district A, where only two white male candidates are competing, and district B, where two black male candidates are competing. Furthermore, suppose these districts are similar to one another in observable characteristics, and the candidates in A and B differ only by race. Note that the Tullock contest function across these two congressional districts is identical, so any difference in spending by l must arise from differences in the weights l assigns to white versus

[20] Note that $\sigma_{\xi_q|PW}$ and σ_{cost} impose the following variance-covariance matrix on the unobservables that enter the preferences of candidates, interest groups and party leadership,

$$\Omega = \begin{bmatrix} \sigma^2_{cost,C}+\sigma^2_\xi & 0 & \sigma^2_\xi & 0 & \sigma^2_\xi & 0 \\ 0 & \sigma^2_{cost,C}+\sigma^2_\xi & 0 & \sigma^2_\xi & 0 & \sigma^2_\xi \\ \sigma^2_\xi & 0 & \sigma^2_{cost,IG}+\sigma^2_\xi & 0 & \sigma^2_\xi & 0 \\ 0 & \sigma^2_\xi & 0 & \sigma^2_{cost,IG}+\sigma^2_\xi & 0 & \sigma^2_\xi \\ \sigma^2_\xi & 0 & \sigma^2_\xi & 0 & \sigma^2_{cost,PL}+\sigma^2_\xi & 0 \\ 0 & \sigma^2_\xi & 0 & \sigma^2_\xi & 0 & \sigma^2_{cost,PL}+\sigma^2_\xi \end{bmatrix} \tag{4.3}$$

Note that the parameter $\sigma_{\xi_q|PW}$ is recovered by the pairwise correlation in campaign participation (and spending) decisions made by players within a congressional district. The variance of the spending levels chosen by player l, along with the dispersion in participation decisions, identifies $\sigma_{cost,l}$.

black candidates. Now, consider a district M, identical to A and B, but with one white and one black candidate. Given that we can infer $\beta_{q,l}$ from comparing A and B, differences in spending and voting in M versus A (and B) can be used to recover $\beta_{q,v}$. I demonstrate these arguments formally in Appendix B.

Estimation: Not all districts experience positive levels of spending, leading to variation along the extensive margin of spending. To address this, I construct a penalized likelihood to estimate the model. The principles of the identification strategy remain valid. For congressional district d and shocks ξ_d, let the set of equilibrium participation profiles be given by $\mathcal{E}_d = \{a_e^* : e = 1, 2, \ldots, E_d\}$, where E_d is the number of pure strategy equilibria in race d. The corresponding payoff that player l, aligned with i, receives under equilibrium a_e^* is given by $V_{i,d,l}^{II}(a_e^*; W_d, \xi d)$. Define the sum of payoffs all players recieve under equilibrium a^* as $\bar{V}(a_e^*, Wd, \xi_d) = \sum_i \sum_l V_{i,d,l}^{II}$. Then, the probability that equilibrium participation profile A_d is played is given by

$$Prob\left[A_d; \{Q_{i,d}, p_{i,d}\}_{i\in\{1,2\}}, X_d\right] = \mathbb{E}_{\xi_d}\left[\overbrace{\frac{\exp\left(\bar{V}\left(A_d, W_d, \xi_d\right)\right)}{\sum_{e=1}^{E_d}\exp\left(\bar{V}\left(a_k^*, W_d, \xi_d\right)\right)}}^{\text{Equilibrium Selection}} \cdot \underbrace{\mathbb{1}\left\{A_d \in \mathcal{E}_d\right\}}_{A_d \text{ is an Equilibrium}}\right], \tag{4.4}$$

where the operator $\mathbf{E}_{\xi_d}[.]$ is the expectation operator with respect ξ_d. The indicator function $\mathbb{1}\{A_d \in \mathcal{E}_d\}$ checks if A_d is indeed an equilibrium given race d characteristics and shocks ξ_d. The multinomial logistic function is the probabilistic equilibrium selection rule that falls under Assumption 2.2. The probability that electoral outcome is $Y_d = 1$ given the participation profile is A_d is given by $Prob\left[1 \text{ wins} \middle| \left\{s^*(A_d), Q_{i,d}, p_{i,d}, \xi_{q,i,d}\right\}_{i\in\{1,2\}}; X_d\right]$ where $s^*(A_d)$ is the equilibrium spending level given participation profile A_d. Then the overall probability of observing (Y_d, A_d) is given by:

$$\begin{aligned}\ell\Big(\theta; Y_d, A_d, Q_d, p_d, X_d\Big) = \mathbb{E}_{\xi_d}\Bigg[&\frac{\exp\left(\bar{V}\left(A_d, W_d, \xi_d\right)\right)}{\sum_{e=1}^{E_d}\exp\left(\bar{V}\left(a_e^*, W_d, \xi_d\right)\right)} \cdot \mathbb{1}\left\{A_d \in \mathcal{E}_d\right\} \\ &\times\left(\sum_{y\in\{0,1\}} \mathbb{1}\left\{Y_d = y\right\} \cdot Prob\left[y \middle| \left\{s^*(A_d), Q_{i,d}, p_{i,d}, \xi_{q,i,d}\right\}_{i\in\{1,2\}}; X_d\right]\right)\Bigg],\end{aligned} \tag{4.5}$$

where $y = 1$ indicates $i = 1$ wins the election. In addition to election result and campaign participation decisions, I use a set of moments which are correlation of spending levels of player l align with i and candidate characteristics $Q_{i,d}$. Given this the penalized likelihood is given by

$$\begin{aligned}\ell\ell\Big(\theta; Y, A, Q, p, X, S\Big) = &\frac{1}{D}\sum_d \log(\ell\left(\theta; Y_d, A_d, Q_d, p_d, X_d\right)) \\ &- \frac{\rho}{D}\left(\sum_{k=1}^{K_Q}\sum_{i=1}^{2}\sum_{l\in\{C,IG,PL\}}\left\{1 - \frac{\sum_{d=1}^{D}\mathbb{E}_{\xi_d}\left[s_{i,d,l}(A_d, Q_d, p_d, X_d, \xi_d)\cdot Q_{i,d,k}\right]}{\sum_{d=1}^{D} S_{i,d,l}\cdot Q_{i,d,k}}\right\}^2\right),\end{aligned} \tag{4.6}$$

where K_Q is the number of characteristics of candidates that I observe, $S_{i,d,l}$ is observed spending level, $s_{i,d,l}(A_d, Q_d, p_d, X_d, \xi_d)$ is the equilibrium spending level given observed campaign participation profile (A_d), and ρ is the penalization hyper parameter. The expectation operators are calculated using importance sampling. Table A3 reports the results of Monte-Carlo experiments using sample sizes of 500, 1500, and 3000. For each parameter the MSE reduces as the sample size increases.

4.2 Entry and Primary Stage

Identification: In order to estimate the EP stage, I use primary winner valence characteristics ($Q^W_{d,P}$), primary winner platform positions ($p^W_{d,P}$), congressional district d specific average primary winner valence shocks ($\xi^W_{d,P} = X'_d\beta_\xi$) where β_ξ is estimated in GE stage, and mass of primary race contestants ($E_{d,P}$) as outcome variables. The set of parameters that need to be estimated for this stage are given by

$$\Theta^{EP} \equiv \left(\left\{\delta_{q,P}, \mu_{p,P}, \sigma_{p,P}\right\}_{P\in\{R,D\}}, \delta_I, \mu_\xi, \sigma_\xi\right), \tag{4.7}$$

where $\delta_{j,q,P}$ is the weight that the median voter of party P in district d puts on candidate characteristic $Q_{j,q}$; $\mu_{p,P}$ is the mean platform position of the potential candidate distribution of party P; $\sigma_{p,P}$ is the standard deviation of the platform position distribution of party P; $\delta_{j,I}$ is the coefficient of $X_{j,d,Prim}$ that explains the variation in the median voter's preferred platform position due to the variation in $X_{j,d,Prim}$; μ_ξ is the mean of the valence distribution of congressional candidates; and σ_ξ is the standard deviation of the congressional candidate valence distribution. I assume that the mass of potential contestants λ is known and is given by the maximum number of contestants observed in a primary race, which is 19. This assumption is stated in Assumption 4.2.

Assumption 4.2 *Mass of potential candidates, λ, is known and is equal to the maximum number of candidates contesting across all primaries.*

Moreover, the econometrician knows the underlying candidate characteristic distribution, which is defined in Assumption 2.3. I recover these distributions from the Census and ACS. This assumption is stated in Assumption 4.3.

Assumption 4.3 *The econometrician knows $N(\mu_{cont}, \Sigma_{cont})$ and $Prob\left[Q_{l,disc} = x\right] = q_{x,l,disc}$ for $l = 1, 2, \ldots, K_{disc}$ such that $x \in \boldsymbol{Q}_{l,disc}$ and $\sum_{x\in\boldsymbol{Q}_{l,disc}} q_{x,l,disc} = 1$.*

The parameter δ_I is identified by correlation of primary winner platform position ($p^W_{d,P}$) and the district observable characteristic X_d. This is demonstrated in Figure 2a. The parameter $\mu_{p,P}$ has a one-to-one relationship with the mean of primary winner platform position ($\mu^W_{p,P} = \mathbb{E}\left[p^W_{d,P}\right]$), therefore the mean platform position of primary winners for a specific

party identifies $\mu_{p,P}$. Figure 2b demonstrates this relationship.[21]

The standard deviation of potential candidate platform positions, $\sigma_{p,P}$, also has a one-to-one relationship with the standard deviation of primary winner platform positions ($\sigma^W_{p,P} = V\left[p^W_{d,P}\right]$). Higher is this dispersion in party P's potential candidates's platform positions, the more dispersed will be platform positions of primary winners of party P. This is demonstrated in Figure 2c. Mean valence shock (μ_ξ) of potential candidates is identified by the mean of valence shock of primary winners ($\mu^W_{\xi,P} = \mathbb{E}\left[\xi^W_{d,P}\right]$). Note that the primary winner's mean valence shock is recovered from the GE estimation stage.

The weights primary voters place on candidate characteristics, $\delta_{q,P}$, are identified by the proportion of primary winners who share those characteristics in party P. This is demonstrated in Figure 3a. It is important to note that this identification is not possible if the expected payoffs that candidates receive from winning the primaries are unknown. If these payoffs were unknown, then the share of primary winners from a social group would only identify the contour of the payoff from winning the primary and the likelihood of winning, without isolating these two factors. However, due to the GE stage parameter estimates, we can compute these payoffs for all possible (observed and unobserved) pairs of primary winners, so the proportion of primary winners with some characteristic x is sufficient to identify the weights primary voters associate with those characteristics.

Finally, the standard deviation of valence shocks is recovered by the mass of primary entrants ($\bar{E}_P$) and not by the standard deviation in primary winner valence shocks ($\sigma^W_{\xi,P} = \sqrt{V\left[\xi^W_{d,P}\right]}$). The moment $\sigma^W_{\xi,P}$ has a non-monotonic relationship with σ_ξ and therefore fails to provide pointwise identification for this parameter. As the dispersion of valence shocks among potential candidates increases, there is a direct effect that leads to greater dispersion in the valence of primary winners. Simultaneously, increased dispersion of valence means that candidates with lower valence have a reduced likelihood of winning the primary, as the mass of candidates with extraordinary valence is higher. This discourages them from entering and contesting in the primary race. Initially, the direct effect of increased valence dispersion dominates, resulting in higher dispersion of valence shocks among the primary winners. However, as the dispersion in valence shocks continues to rise, the indirect effect—where lower-valence candidates opt out due to a diminished likelihood of winning—becomes stronger and eventually prevails, leading to a reduction in the dispersion of valence shocks among primary winners. Therefore, $\sigma^W_{\xi,P}$ fails to be a suitable moment for pointwise identification. This relationship is demonstrated in Figure 3c.

[21] It is important to note that this shift in mean of winner platform position can also be introduced by the constant term in vector of district observable characteristics. Due to this, I calibrate constant term's coefficient to the value 0. Moreover, the observable characteristics are standardized that ensures mean primary voter ideology is 0. Therefore,$\mu_{p,P}$ are relative to the average districts' preferred ideology across the two parties.

When potential candidates become more dispersed in valence, candidates with relatively lower valence choose not to contest in the primaries, as they now must compete with a higher mass of candidates with extraordinarily high valence to win the primary. This reduces the expected payoff for candidates with lower valence from contesting, resulting in candidates with relatively lower valence opting out of primaries and, in the process, reducing the mass of entrants. This selection mechanism introduces a monotonic relationship between the mass of entrants, $\bar{E}_P$, and the dispersion in valence among potential candidates, σ_ξ. As a result, the average mass of entrants in the primary race identifies the standard deviation of valence shocks among potential candidates. This is demonstrated in Figure 3f.

Estimation: In terms of outcomes, for $P \in \{R, D\}$ we observe primary winner characteristics $Q^W_{P,d,k}$ for $k = 1, 2, \ldots, K_Q$, primary winner's platform positions $p^W_{P,d}$, mass of enterants in primary $E_{P,d}$. In addition to this, we also observe the mean valence of primary winners recovered from GE stage given by $\bar{\xi}^W_d = X'_d \hat{\beta}_\xi$, where $\hat{\beta}_\xi$ is the estimate of β_ξ from the GE state. I use the Method of Simulated Moments (McFadden, 1989) to estimate the model. The model predictions are simulated by solving the model for a set of $R = 200$ potential candidates in each party. Note that setting $R = 200$ gives me a set of $R^2 = 40,000$ potential pairs of primary winners as any candidate can win the primary with positive probability if they enter. From these simulations, I obtain the model's predicted versions defined as

$$\begin{aligned}
\hat{Q}^W_P(\theta, X^{Prim}_d) =& \mathbb{E}_{F_P}\left[Q \cdot \Pi^{ex-ante}_P(Q, p, \xi, X^{Prim}_d) | X^{Prim}_d\right] \quad \text{for } P \in \{R, D\}, \\
\hat{p}^W_P(\theta, X^{Prim}_d) =& \mathbb{E}_{F_P}\left[p \cdot \Pi^{ex-ante}_P(Q, p, \xi, X^{Prim}_d) | X^{Prim}_d\right] \quad \text{for } P \in \{R, D\}, \\
\hat{E}_P(\theta, X^{Prim}_d) =& \lambda \cdot \mathbb{E}_{F_P}\left[\pi^{ex-ante}_P(Q, p, \xi, X^{Prim}_d) | X^{Prim}_d\right] \quad \text{for } P \in \{R, D\}, \\
\hat{\xi}^W(\theta, X^{Prim}_d) =& \frac{1}{2} \sum_{P \in \{R,D\}} \mathbb{E}_{F_P}\left[\xi \cdot \Pi^{ex-ante}_P(Q, p, \xi, X^{Prim}_d) | X^{Prim}_d\right],
\end{aligned} \tag{4.8}$$

where $\mathbb{E}_{F_P}$ is the expectation operator with respect the potential candidate distribution, F_P, of party P. Moreover, $\Pi^{ex-ante}_P$ is the ex-ante (prior to the order of primaries is realized) equilibrium winning density of candidate with type (Q, p, ξ). This density can be calculated by using the Propositions 2.2 and 2.3. Finally, $\pi^{ex-ante}_P$ is the ex-ante equilibrium entry function of type (Q, p, ξ) from party P. This can also be calculated using the Propositions 2.2 and 2.3. Define $Y_d = \left(\{Q^W_{P,d,k}, p^W_{P,d}, E_{P,d}\}_{P \in \{R,D\}}, \bar{\xi}^W_d\right)$, and $Y^{squared}_d = ((p^W_{R,d})^2, (p^W_{D,d})^2)$. The model predictions are defined as

$$y(\theta, X^{Prim}_d) = \left(\{\hat{Q}^W_P(\theta, X^{Prim}_d), \hat{p}^W_P(\theta, X^{Prim}_d), \hat{E}_P(\theta, X^{Prim}_d)\}_{P \in \{R,D\}}, \hat{\xi}^W(\theta, X^{Prim}_d)\right), \tag{4.9}$$

$$y_{square} = \left(\mathbb{E}_{F_R}\left[p^2 \cdot \Pi^{ex-ante}_R(Q, p, \xi, X^{Prim}_d) | X^{Prim}_d\right], \mathbb{E}_{F_D}\left[p^2 \cdot \Pi^{ex-ante}_D(Q, p, \xi, X^{Prim}_d) | X^{Prim}_d\right]\right). \tag{4.10}$$

Then the estimator is defined as

$$J(\theta) = \frac{1}{D} \sum_{l=1}^{2K_Q+5} \frac{(Y_{d,l} - y_l(\theta, X_d))^2}{V(Y_{d,l})} + \frac{1}{D} \sum_{l=1}^{2} \frac{\left(Y^{squared}_{d,l} - y^{squared}_l(\theta, X_d)\right)^2}{V\left(Y^{squared}_{d,l}\right)}. \tag{4.11}$$

The above construction is important because it utilizes information not only from conditional means that are needed for identification of δ_q, δ_I, μ_ξ, $\mu_{R,p}$, $\mu_{D,p}$ and σ_ξ but also the spreads captured by Y_d^{square} and $y^{square}(\theta, X_d)$ which is needed for identification of $\sigma_{p,P}$.

5 Results

5.1 General Election Stage Results

Table 4 reports the estimates for the General Election stage. Specifically, it reports estimates of spending effectiveness (β), the ideal policy position of the representative interest group (I_{IG}), the weight party leadership places on policy preferences ($\omega_{I,PL}$), the standard deviation of cost shocks ($\sigma_{cost,l}$ for $l \in \{C, IG, PL\}$), the standard deviation of unobserved valence shocks (σ_ξ), candidate preferences over valence characteristics ($\beta_{q,C}$), interest group preferences over valence characteristics ($\beta_{q,IG}$), party leadership preferences over valence characteristics ($\beta_{q,PL}$), and voter preferences over valence characteristics ($\beta_{q,V}$). Median voter ideology and conditional mean valence estimates are provided in Table A2.

The estimate for effectiveness of spending 1000 USD is 0.0477 (0.009).[22] The ideal policy position for representative IG is estimated to be 0.293 (0.205), which is not significantly different from 0, indicating that the representative interest groups take moderate policy position. The weight party leadership puts on policy preferences is estimated to be 1.26, which is significantly greater than 0. Variance in candidate preferences is much higher than that of IG and party leadership.

Now we discuss the main parameters of this paper, which are the preferences of candidates, interest groups, party leadership, and voters regarding the race and gender of the candidates.[23] The candidate preferences relate to pure election aversion, which is discussed in the literature on the underrepresentation of women (Anzia and Berry, 2011; Ashworth et al., 2024). It is important to note that Hispanic Americans and White Americans value winning a congressional district more than African Americans or Americans belonging to other races. Moreover, males have a higher value for winning a seat in the U.S. Congress than their female counterparts. Therefore, candidates from racially underrepresented groups (specifically Black and other races) and female candidates have a lower political ambition than White Americans and males, respectively. This confirms that these groups are relatively averse to running and winning office. However, whether this pure election aversion

[22] This coefficient translates to average marginal effects of 4.8×10^{-4} (4.7×10^{-5}) for candidates, 0.00864712 (0.000997281) for interest groups, and 0.0015 (0.00015) for party leadership. The average marginal effect estimates are calculated at the observed level of spending. Since the model has decreasing returns to scale for spending, and candidates spend more, we observe higher marginal effects of spending for IG and PL.

[23] The units for the estimates of C, IG, and PL preference parameters is natural log of 1000 USD.

translates into a leading cause of underrepresentation will be confirmed in counterfactuals.

Estimates of the parameters governing interest group preferences indicate that, among candidates with similar valence, platform positions, and winning probabilities, interest groups have a higher preference for Black Americans over other racial groups. This is evident from the estimates: $\beta_{q,IG,Black} = 4.3 > \beta_{q,IG,White} = 1.94$, $\beta_{q,IG,Hispanic} = 2.68$, $\beta_{q,IG,Others} = 0.606$. Moreover, estimates also show that interest groups have a higher preference for female candidates over male candidates, as indicated by $\beta_{q,IG,Male} = -3.56$ (0.335).

Party leadership preferences estimates show that Black candidates are preferred over White candidates and candidates from Other races, both at the 90% level of confidence. Preference for Hispanic candidates is not significantly different from both either Black candidates nor from White candidates. There is also no sign of gender-based discrimination for party leadership. Hassell and Visalvanich (2019) have analyzed party preferences over candidate race and gender during the primaries. They did not find evidence of discrimination on race, but found that Democratic Party preferences to be biased towards White Women. Overall, our results agree with theirs since we do not focus on party-specific preferences.[24]

General election contest function estimates, denoted by $\beta_{q,V}$, indicate that Black candidates ($\beta_{q,V,Black} = -1.66$, standard error 0.265) and candidates from other races ($\beta_{q,V,Other} = -2.13$, standard error 0.31) are less preferred than White candidates.[25] On the other hand, Hispanic candidates ($\beta_{q,V,Hispanic} = 0.71$, standard error 0.291) are slightly more preferred to White candidates. Furthermore, general election voters exhibit a preference for male candidates over female candidates.

Table 5 reports the model fit for the GE stage. The model accurately captures candidate spending levels across party affiliation, race, and gender. Predictions for interest group (IG) spending by party affiliation and gender closely match the observed data. The model overestimates interest group spending in support of Black candidates. The predicted spending decisions of party leadership are also closely aligned with the observed values across party affiliation, race, and gender. Overall, the model replicates the spending behaviors of the various players with a high degree of accuracy.

In addition to spending decisions, the model fit for the voting decisions is also assessed. Note that in the estimation of the GE stage, I use only the win or lose outcome of the candidate. Due to this, it is possible that the model's estimated parameters may not differentiate competitive races from lopsided races, which may bias the estimates. The comparison of the model's win probabilities with the observed vote shares shows that this is not the case.

[24] The model has 46 parameters at this stage; allowing for preferences to be party-specific will increase the number of parameters greatly, and we may not have sufficient statistical power.

[25] Note that $\beta_{q,V,White}$ is normalized to 0.

Note Black candidates on average received 33% of the two-party vote, and the average winning probability of the Democratic Black candidates is also 33%. The rest of the categories are more on the margin in the data, and the model's win probabilities also predict the same.

5.2 Entry and Primary Stage Results

Table 6 reports the parameter estimates for the EP stage of the model. For inference, I follow Newey and McFadden (1994) to calculate the standard errors for a two-step GMM estimation. The table reports the estimates for the potential candidate distribution parameters, primary voter preferences over candidate characteristics by party, and median voter ideology estimates, which also include institutional controls.[26]

The mean valence of potential candidates is $\mu_\xi = -5.51$, which is much lower than the average valence of primary winners (≈ 0.00). The standard deviation of the valence distribution, σ_ξ, is 3.41, indicating significant variation in the valence of potential candidates. The estimated means of the platform position distribution for Republicans and Democrats are not significantly different from the average platform positions of their respective primary winners. Specifically, the estimated mean platform positions of potential Republican and Democrat candidates are $\mu_{p,R} = 1.15$ (0.0613) and $\mu_{p,D} = -0.865$ (0.0721), compared to the platform positions of primary winners, which are 1.00 and 0.70, respectively.

The preferences over racial and gender characteristics of primary voters show evidence of discrimination. In Table 6, I normalize $\delta_{q,P,Other} = 0$ for $P \in \{R, D\}$ and report the remaining racial coefficient estimates. I find that Hispanics are significantly less preferred than other races by Republican primary voters, as $\delta_{q,R,Hisp} = -4.71$ (0.4). However, this does not hold for Democratic primary voters, who are indifferent between Hispanic candidates and other races, $\delta_{q,D,Hisp} = -0.307$ (1.16). Republican primary voters are indifferent between White candidates and candidates from other races, $\delta_{q,R,White} = -0.758$ (0.0848), while Democratic primary voters strictly prefer White candidates over candidates from other races, $\delta_{q,D,White} = 3.44$ (0.325). For Republican primary voters, Black candidates are less preferred than other race candidates and also less preferred than White candidates, $\delta_{q,R,Black} = -3.8$ (0.275). For Democratic primary voters, Black candidates are preferred over candidates from other races but still less preferred than White candidates, $\delta_{q,D,Black} = 2.09$ (0.735). Primary voters from both parties prefer male candidates over female candidates, with $\delta_{q,R,Male} = 3.21$ (0.388) and $\delta_{q,D,Male} = 2.27$ (1.32).

[26] Top-two primaries are incorporated in these estimates. For simulating the top-two primaries, I assume that the potential candidate distribution is a 1:1 mixture of Republican and Democrat potential candidate distributions. Moreover, the top-two voter preferences are also assumed to be the average of party-specific voter preferences.

Primary voter ideal point estimates are also provided. Districts with a higher proportion of Whites, Blacks, and males are more conservative. Over time, there is a trend of primary voters becoming more conservative. Districts with higher median household income and a lower proportion of college graduates are more left-leaning. Republican primary voters in districts with a higher lagged Democratic presidential vote share are more left-leaning, whereas a weaker relationship exists for Democratic primary voters. On average, Democratic primary voters voting in open primaries are more conservative than those voting in closed primaries. Semi-closed Democratic primary voters are more left-leaning. For Republicans, open primary voters are more conservative than closed and semi-closed primary voters.

Table 7 reports the model fit for the Entry and Primary stage. The predicted average probability of White candidates winning primaries, either through Republican or Democratic primaries, is very close to the observed values. The same holds for minority races, such as Hispanics, Whites, and Other races, in Republican primaries. For Democratic primaries, except for Hispanic winners, the model's predicted probability of Black or Other race candidates winning the primary is also quite close to the observed values. However, I overestimate the average probability of a Hispanic candidate winning the primary for Democratic candidates. Overall, the model provides a good fit for winning outcomes based on the race of candidates. The same applies to winning outcomes based on the gender of candidates. The predicted average platform positions of primary winners are also quite close to the observed values. Additionally, the predicted average number of Republican and Democratic candidates is close to the observed number.

6 Decomposing the sources of underrepresentation

In this section, I decompose the extent to which individual factors, such as the preferences of candidates, voters, interest groups, and party leadership, contribute to underrepresentation. To do this, I adjust the parameters of each player, making them indifferent to the gender and race of candidates, one at a time. For each case of indifferent preferences, I solve for the equilibrium entry and spending decisions, along with contest function outcomes for all congressional races. Using these equilibrium outcomes, I calculate the share of entrants, primary winners, and general election winners from underrepresented groups. Finally, I report the results.

There are two types of voters: general election voters and primary voters. To make general election voters indifferent across race and gender, I multiply $\beta_{q,V}$ by 0, leaving the preferences of other players (including primary voters) unchanged. To make primary voters indifferent across race and gender, I multiply $\delta_{q,R}$ and $\delta_{q,D}$ by 0 while keeping the prefer-

ences of other players the same. To make the preferences of $l \in \{C, IG, PL\}$ indifferent across race and gender, I replace $\beta_{q,l}$ with the value $\bar{\beta}_{q,l} = \sum_{k\in\{Hisp,Black,White,Other\}} \beta_{q,l,k} Prop_k + (1 - Prop_{Male})\beta_{q,l,Male}$. The term $\bar{\beta}_{q,l}$ represents the average coefficient associated with the race and gender of the candidate, where $Prop_k$ is the average population share of social characteristic k.[27]

For each case, I compute the model's predicted share of entrants, primary winners and general election winners from social group k. Let these quantities be denoted as $Q_k^{E,l}$, $Q_k^{PW,l}$, and $Q_k^{GEW,l}$, where E stands for entrant, PW stands for primary winner, and GEW stands for general election winner. The term $l \in \{Prim, GE, C, IG, PL\}$ indicates whose preferences are being forced to be indifferent across race and gender. The term $k \in \{Hisp, Black, Others, AllMinorities, Female\}$ denotes the social group. Let $Q_k^{E,o}$, $Q_k^{PW,o}$, and $Q_k^{GEW,o}$ denote the shares obtained under the estimated preferences for entrants, primary winners, and general election voters. Then, define the following proportional change in underrepresentation:

$$\begin{aligned}
\Delta\text{Underr}_k^{E,l} &= \frac{(Prop_k - Q_k^{E,l}) - (Prop_k - Q_k^{E,o})}{(Prop_k - Q_k^{E,o})}, \\
\Delta\text{Underr}_k^{PW,l} &= \frac{(Prop_k - Q_k^{PW,l}) - (Prop_k - Q_k^{PW,o})}{(Prop_k - Q_k^{PW,o})}, \\
\Delta\text{Underr}_k^{GEW,l} &= \frac{(Prop_k - Q_k^{GEW,l}) - (Prop_k - Q_k^{GEW,o})}{(Prop_k - Q_k^{GEW,o})},
\end{aligned} \tag{6.1}$$

where the denominators $Prop_k - Q_k^{E,o}$, $Prop_k - Q_k^{PW,o}$, and $Prop_k - Q_k^{GEW,o}$ are interpreted as the share of the population belonging to group k that does not have representation in Entrants, Primary Winners, and General Election winners respectively. Therefore, $\Delta\text{Underr}_k^{PW,l}$ and $\Delta\text{Underr}_k^{PW,l}$ report the proportional change in underrepresentation of social group k.

The results of this exercise are reported in Figures 6 and 7. Let us consider the decrease in underrepresentation when general election voters are made indifferent. Refer to Figure 6. If there were no discrimination by general election voters, underrepresentation of Black candidates could be substantially reduced in the U.S. House. This amounts to a 6.92% reduction in underrepresentation of Black Republican congresspersons and a 39.75% reduction in underrepresentation of Black Democratic congresspersons. This is a result of both an increase in the entry of Black candidates through Democratic primaries (Panel (a) in Figure 6) and a substantial increase in their general election winning probabilities. For Hispanics, there is a slight increase in underrepresentation. The net effect on all racial minorities, however, is much smaller. For Republican general election winners, the net effect is a decrease of 0.63%, while for Democrats, it results in a 6.42% reduction. For female politicians, gen-

[27] This substitution is required instead of replacing $\beta_{q,l}$ with 0 because for $l \in \{C, IG, PL\}$ the whole vector, $\beta_{q,l}$ is identified. For voters, one of the coefficients for race has to be normalized to 0.

eral election voters being indifferent to gender could increase the representation of women among Democrats by 5%, while for Republicans, the increase is much smaller at 0.5%.

If interest groups were indifferent, underrepresentation would increase, particularly for Democratic primary winners compared to Republicans or general election primary winners. This increase occurs because interest groups have an affinity for underrepresented groups.[28] However, this affinity does not necessarily lead to success in the general elections. Therefore, the presence of interest group campaigning has actually improved representation among primary winners by supporting socially underrepresented candidates to contest and win primaries, but this effect does not seem to translate into increased representation in Congress.

The effect of party leadership's taste-based discrimination is negligible. Making them indifferent seems to increase underrepresentation among Republican Hispanic GE winners, Republican Black GE winners, and Democratic Black primary winners, though the magnitude of these changes is quite small.

Now, we discuss the effect of candidate preferences. Recall that candidate preferences refer to pure election aversion, which scholars have identified as a leading cause of underrepresentation (Fox and Lawless, 2004, 2011; Lawless and Fox, 2010; Kanthak and Woon, 2015). If there were no differences in political ambition among candidates by race and gender, entry from underrepresented groups would increase substantially. However, this higher entry does not lead to improved chances of winning the general elections. While there is a substantial decrease in the underrepresentation of Black candidates, the overall reduction in all minority groups is not significant, and for females, the decrease is minimal. This finding suggests that even if the political ambitions of women were higher, they would still be less likely to enter politics due to lower success rates. Therefore, differences in political ambition are not the primary cause of underrepresentation.

Overall, among the general election players, the maximum decrease in underrepresentation in the U.S. House arises from making general election voters indifferent across race and gender of candidates. Now we study the changes in underrepresentation when there is no taste based discrimination in primary voters' preferences. Refer to Figure 7, and note that for most social groups, roughly a 70% reduction in underrepresentation would be observed if primary voters were indifferent across the race and gender of candidates. This substantial decrease in underrepresentation results from two factors.

First, there is a higher entry of candidates from these social groups compared to the baseline case, due to an improvement in primary win densities—in other words, an improve-

[28] Recall from Table 4, Hispanics and Blacks are substantially more valued than other races and Whites. Moreover, interest groups prefer females over males.

ment in their chances of advancing to the general election stage. This change in the pool of contesting candidates and the elimination of discrimination by primary voters both contribute to increasing the representation among primary winners.

At the general election stage, the pool of contesting candidates is substantially more representative than before. This supply-side change in the general election stage essentially leads to the improvement in representation in the U.S. House of Representatives. Moreover, the effect of discrimination by general election players, including the general election voters, is much weaker (except for the underrepresentation of Black candidates) and hence does not succeed in suppressing the improvements in minority representation.

This finding is crucial in identifying the main source of underrepresentation, a topic of debate among scholars studying the political economy of gender and race in the United States. It shows that pure election aversion is not the primary cause of underrepresentation; rather, it is discrimination by primary voters. In fact, primary voter discrimination severely distorts the incentives of candidates from underrepresented groups, leading to lower entry of candidates from these groups and making underrepresentation appear as a consequence of election aversion. Moreover, it highlights that discrimination by general election voters is not the primary concern regarding underrepresentation. Instead, it is the discrimination by primary voters that drives the significant patterns of political selection.

7 Counterfactual Policy Experiments

In this section, I evaluate the effects of three policies that can be used to improve representation. These policies are campaign support for underrepresented groups during general elections, campaign support for underrepresented groups during primary races, and quota for underrepresented groups. To analyze the effects of these policies I consider four measures.

The first measure is the share of individuals from underrepresented groups, defined here by race-gender pairs. Note that white males, who represent an average of 32% of the population, held an average of 73.2% of congressional seats. Therefore, I classify all race-gender pairs excluding white males as underrepresented, and denote this set of pairs by Γ. I define the share of general election winners who are not white males as the share of underrepresented groups among GE winners. This share is similarly calculated for primary winners, which constitutes our second measure. The third measure captures ideological polarization, represented by the absolute difference between the average Republican and Democrat platform positions. Lastly, the fourth measure is the average valence shock, which reflects candidate quality in this framework. Therefore, a decline in average valence shock is interpreted as a decrease in candidate quality.

7.1 Campaign Support Subsidy in General Election

Campaign subsidies are not unfamiliar in electoral competition. For instance, U.S. presidential candidates who meet certain conditions are eligible for campaign funding from the government.[29] Countries like Canada, Sweden, Finland, Israel, Australia, and many others provide subsidies either to parties or individual candidates (Casas-Zamora, 2005). Here, I introduce a subsidy for campaign expenditures to candidates who belong to underrepresented groups.

To implement this policy, I increase the payoff of candidates by the subsidy amount. That is, $\tilde{U}^{sub}_{i,d,C} = U_{i,d,C} + Sub \cdot \mathbb{1}\{Q_i \in \Gamma\}$. The units of $U_{i,d,C}$ are in USD, so simply adding the subsidy does not change the units of the payoffs. The equilibrium is then solved using $\tilde{U}_{i,d,I}$. This adjustment has two effects: first, spending levels in the General Election increase on average for underrepresented groups. Second, there is higher entry from underrepresented groups, as the subsidy directly targets their political ambition. Recall from Figure 6 that addressing differentials in political ambition increases political entry but does not necessarily translate into better representation for underrepresented groups.

Unfortunately, since this support is made available at the General Election stage, the higher spending levels and increased entries do not lead to substantial changes in representation. The success rate of candidates in the primaries remains unaltered, and the pool of candidates in the General Election stage does not change significantly enough to have meaningful effects, resulting in only slight increases in representation. Figure 6 plots the results of this exercise. Note that the x-axis in both panels is in natural logs of '000 USD. Even if campaign support is increased to enormous amounts, representation will not substantially change.

7.2 Campaign Support Subsidy in Primary Elections

Although I do not explicitly model campaign spending in the primaries, the framework allows for examining the potential impact of such a subsidy. To do this, I first calculate the effectiveness of campaign spending in the primaries, using estimates from the literature. Cox (2022) find that the ratio of the effectiveness of campaign spending in general elections to primaries is 0.61 for Republicans and 0.932 for Democrats. Based on these ratios, the effectiveness of campaign spending in my setting is 0.0291 for Republicans and 0.044 for

[29] For details on public funding of presidential elections, see FEC webpage

Democrats. The large contest function under this policy is then given by:

$$\Pi^W = \frac{\left(e^{Q_i'\delta_q - \left(p_i - X_{d,Prim}'\delta_I\right)^2 + \xi_i} + \beta^{Prim,P}\sqrt{Sub \cdot \mathbb{1}\{Q_i \in \Gamma\}}\right)}{\int_{Q,p,\xi}\left(e^{Q'\delta_q - \left(p - X_{d,Prim}'\delta_I\right)^2 + \xi} + \beta^{Prim,P}\sqrt{Sub \cdot \mathbb{1}\{Q_i \in \Gamma\}}\right) \cdot \pi^* \cdot dF_R(Q,p,\xi)}, \tag{7.1}$$

where π^* is the new equilibrium entry decision. The equilibrium is then solved using this contest function, and the results are plotted in Figure 9. To completely eliminate underrepresentation, a subsidy of 59 million would be required for underrepresented groups. This corresponds to 11 natural logs of 1000 USD in the figure, which is an unreasonably large amount. However, a more reasonable subsidy of 150,000 USD for campaign support to underrepresented groups can significantly improve representation (corresponding to 5 natural logs of 1000 USD). At this subsidy level, representation increases by 177% for Republicans and 30% for Democrats, with an overall improvement of 94% from the base value.

This policy also leads to a decrease in candidate quality, which becomes substantial for higher subsidy amounts. For the recommended subsidy amount of 150,000 USD, the decrease is 16% of the estimated standard deviation of candidate quality. If the campaign support is provided selectively, either by PACs/Super-PACs or committee transfers, the decline in candidate quality can be further minimized, as these groups may engage in screening candidates based on their unobserved valence.

7.3 Quota for Underrepresented Groups

Quota or seat reservations is a policy implemented in many countries. Clayton (2021) pointed out that more than 130 countries have modified their constitution, electoral laws, or party rules to reserve a set of seats or spots for female legislators or candidates. Examples of such countries include Sweden, the United Arab Emirates, Mexico, and Rwanda. Moreover, countries such as India not only reserve seats/constituencies for female candidates (Desai et al., 2024) but also for groups that fall under the category of economically, politically, and socially underprivileged.

The effectiveness of quotas varies from nation to nation depending on the type of quota (candidate quota or seat quota), the electoral system (proportional ranking or first-past-the-post), and the implementation of the quota.[30] I analyze reserved seat quotas in the context of U.S. House elections. In this quota system, a share of seats/congressional districts are

[30] Rosen (2017) found that initiatives by political parties to enforce quotas for candidates are substantially more effective in developed countries, while a constitutional or electoral law mandate that reserves seats was found to be significant in least developed countries. In the study, no developed countries had reserved seat quotas.

reserved for an underrepresented group, where only candidates who belong to that group can enter and compete in elections.

I will execute a set of counterfactuals and calculate the four measures discussed previously. For each counterfactual, I will reserve a share of congressional districts, S_{Quota}. These reserved congressional districts are randomly assigned to a particular underrepresented race-gender pair, with the assignment probability proportional to the share of that race-gender pair in the population of underrepresented groups. In the congressional districts assigned to a particular race-gender pair, only candidates from that race-gender pair can enter and compete in the elections. Specifically, if the pair Black-Females has a population share given by $Prop_{\text{Black-Female}}$, then the share of reserved seats for Black-Females is given by $S_{Quota} \cdot \frac{Prop_{\text{Black-Female}}}{\sum_{j \in \Gamma} Prop_j}$, where $Prop_j$ is the population share of race-gender pair j and Γ is the set of underrepresented race-gender pairs. In the congressional districts assigned to Black-Female candidates, only Black-Female candidates are allowed to enter and compete in the elections. The remaining $1 - S_{Quota}$ congressional districts are open for all race-gender pairs to compete.

The counter-factual experiments are given in Figure 10. In panels (a) and (b), blue refers to Democrats and red refers to Republicans. Panel (a) shows the results for primary winners, and panel (b) shows the results for GE winners. Note that the share of primaries and seats won by underrepresented groups lies above the 45-degree line (the black dashed line), indicating that achieving a 68% share of seats for minorities requires a quota of less than 68%. In this context, the required quota falls between 50% and 60%. Moreover, a 20% quota leads to significant improvements in representation in the U.S. House. For Democrats, there is a 37% increase in representation, while for Republicans, this results in a 101% improvement. These gains are accompanied by an increase in polarization within Congress. The results show that the distance between party average platform positions would increase; however, these shifts are only about 1-2% of baseline polarization. Whether this increase will translate into substantial changes in roll call voting is complex to address in this study (Canen et al., 2020, 2021; Poole and Rosenthal, 1985). There are also signs of improvement in the average quality of candidates, though when compared to the estimated population standard deviation of valence ($\sigma_\xi \approx 3.41$), the increase is not substantial.[31]

8 Conclusion

I propose a tractable model of political entry, voter discrimination, and campaign spending. The model differentiates between the discrimination faced by politicians in the primaries and that in the general elections. It also accounts for discrimination by interest groups and

[31] Due to computational constraints, standard errors of these measures have not been calculated.

party leadership. The model does not place restrictions on the number of competing candidates and succeeds in capturing the effects of candidate pool size on political entry decisions. Even though the model possesses a rich preference specification, it still exhibits a unique Nash Equilibrium in the entry stage. Moreover, the model allows for estimation by providing predictions that can be matched with data.

The discrimination faced by underrepresented groups in the primaries is the main driver of their underrepresentation in the U.S. House. Estimates of model parameters reveal that, although general election voters discriminate against these groups and these groups have lower political ambition, these factors contribute minimally to overall underrepresentation. This finding contributes to the Political Economy of Gender literature, which has analyzed causes of underrepresentation (Fox and Lawless, 2004; Kanthak and Woon, 2015; Anzia and Berry, 2011; Ashworth et al., 2024), as well as to the Political Economy of Race literature that studies representation of racial minorities (Trebbi et al., 2008; Beach et al., 2018; Ricca and Trebbi, 2022; Trounstine and Valdini, 2008; Trounstine, 2010), which has focused on subnational U.S. politics by leveraging rich institutional and demographic variation at local government level.

The paper uncovers two policies that can improve minority representation. These include campaign subsidies during primary elections for candidates from underrepresented groups and the reservation of seats for these groups, with the former being a more feasible option in the U.S. context. However, other policies may also prove to be effective that can influence primary voters. One reason for voter discrimination among primary voters might be lower participation of minorities. While previous studies have analyzed voter turnout in general elections due to the Voting Rights Act (Ang, 2019) or Shelby County v. Holder (2013) (Billings et al., 2024), little is known about turnout in primaries. Higher minority participation in primaries could improve representation without the need for reservations or campaign subsidies.

9 Figures

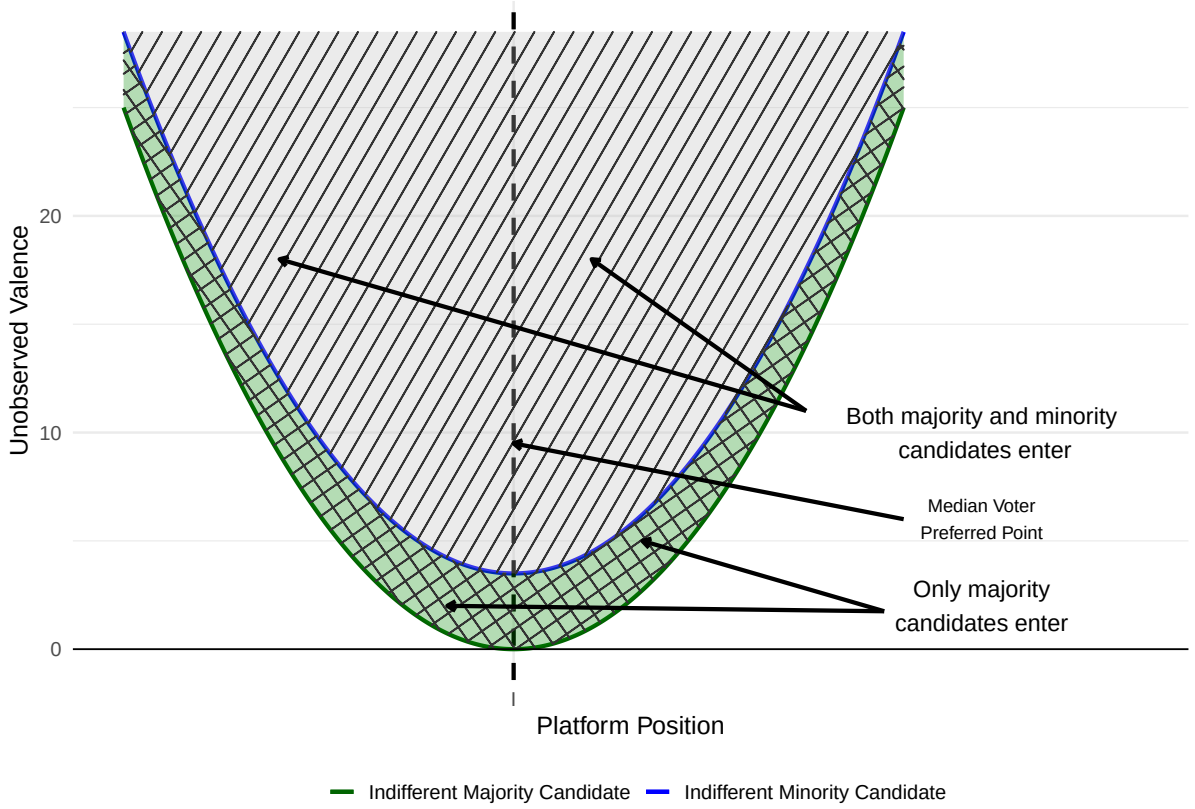
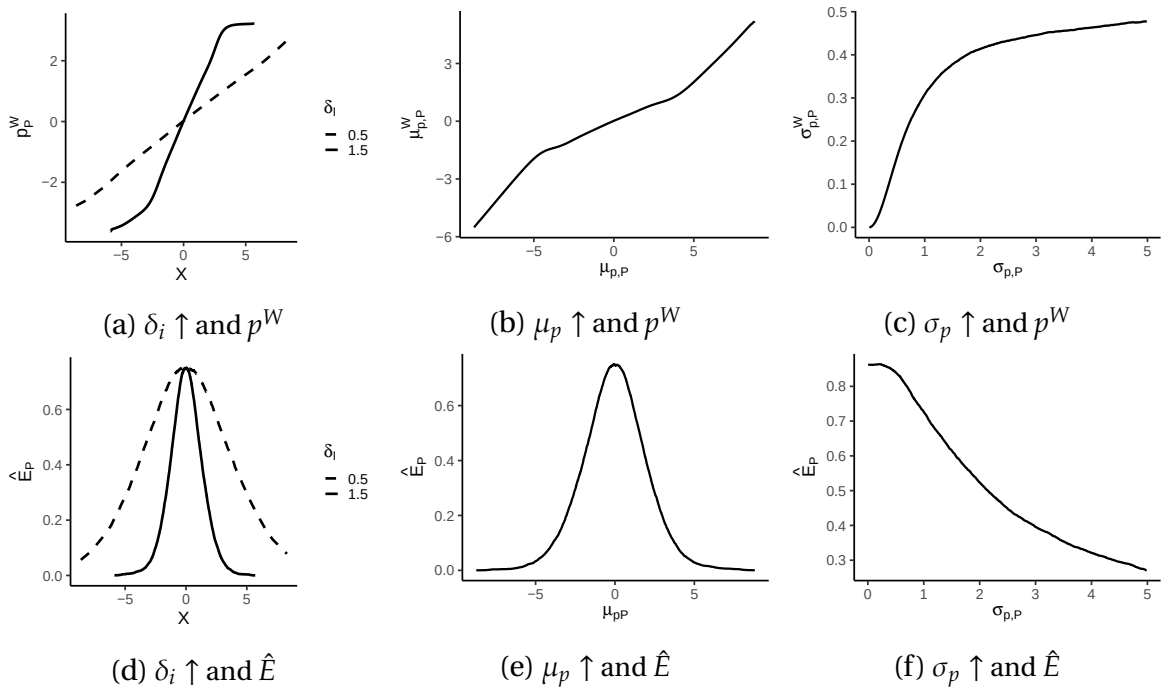


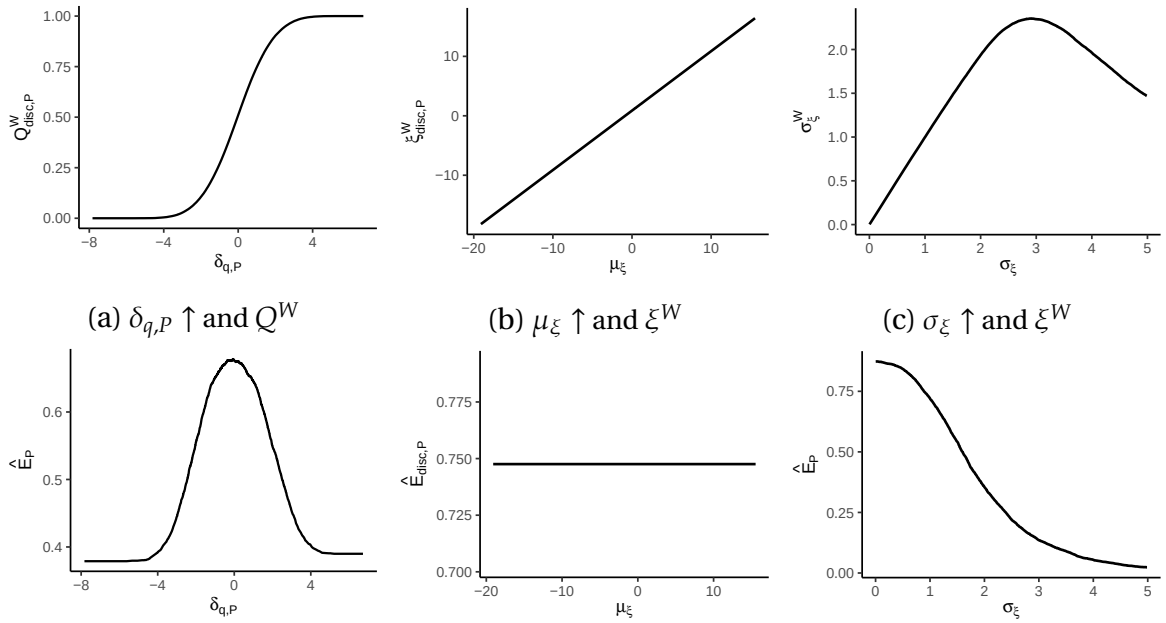
Figure 1: This Figure shows how model accommodates selection on valence and platform position given there is voter discrimination and election aversion. For this example, I assume that candidates belong to two social groups, majority and minority. For simplicity, I assume: (1) There is only one primary. (2) Minority candidate receive V_L if they win the election. (3) Majority candidate receive V_H if they win the election. (4) There is voter discrimination against minority candidates, relative taste based discrimination, $\delta_{Min} < 0$. Then the set of candidates from majority group who choose to enter is given by $\bar{E}^{Maj} = \{(p,\xi) : \xi \geq \log(\frac{\kappa \cdot A}{V_H}) + (p-I)^2\}$ and for minority this set is given by $\bar{E}^{Min} = \{(p,\xi) : \xi \geq \log(\frac{\kappa \cdot A}{V_L}) - \delta_{Min} + (p-I)^2\}$. Note that $\bar{E}^{Min} \subset \bar{E}^{Maj}$. Moreover, also note that the set $\bar{E}^{Maj}/\bar{E}^{Min}$, where only the majority candidates enter, is more distant from median voter platform position and posses lower valence than minority candidates who choose to enter.



(a) $\delta_i \uparrow$ and p^W (b) $\mu_p \uparrow$ and p^W (c) $\sigma_p \uparrow$ and p^W

(d) $\delta_i \uparrow$ and $\hat{E}$ (e) $\mu_p \uparrow$ and $\hat{E}$ (f) $\sigma_p \uparrow$ and $\hat{E}$

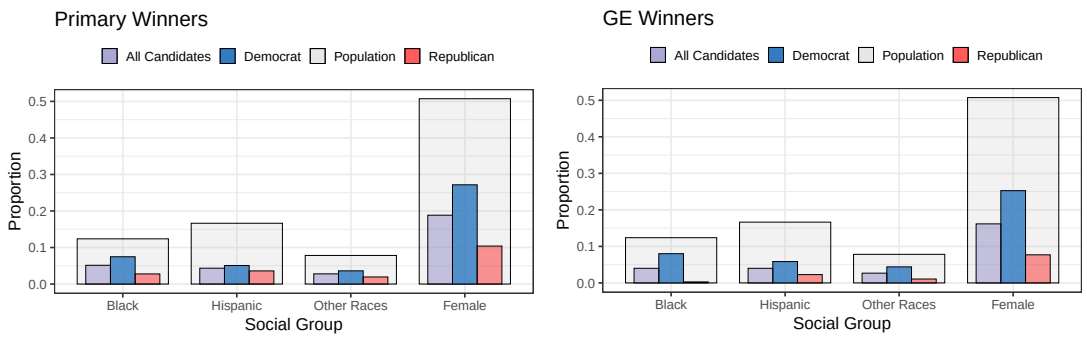
Figure 2: This Figure shows how model predictions change when one changes the parameters δ_I, μ_p, and σ_p. To study this, all potential candidate's expected GE payoff is assumed to be 1 so that we may analyze the changes in the outcomes keeping future payoffs constant. Note that the GE payoff is estimated separately in the GE state. The baseline parameter values are $\mu_{p,P} = 0$ and $\sigma_{p,P} = 1$ for $P \in \{R, D\}$, $\delta_I = 0.5$, $\delta_{q,P} = 0$ for $P \in \{R, D\}$, $\mu_\xi = 0$, and $\sigma_\xi = 0$. Figure 2a shows how the correlation of expected policy position of a primary winner and congressional district characteristic X changes as δ_I is increased. Figure 2d shows how the relationship between the equilibrium mass of entrants and congressional district X changes as δ_I is increased. Figure 2b shows how the mean platform position of primary winners from party P changes as $\mu_{p,P}$ varies. Figure 2e shows how the equilibrium mass of entrants from party P changes as $\mu_{p,P}$ varies. Figure 2c shows how the standard deviation/dispersion of primary winner platform positions changes as $\sigma_{p,P}$ varies for party P. Figure 2f shows how the equilibrium mass of entrants from party P changes as $\sigma_{p,P}$ varies.



(a) $\delta_{q,P} \uparrow$ and Q^W (b) $\mu_\xi \uparrow$ and ξ^W (c) $\sigma_\xi \uparrow$ and ξ^W

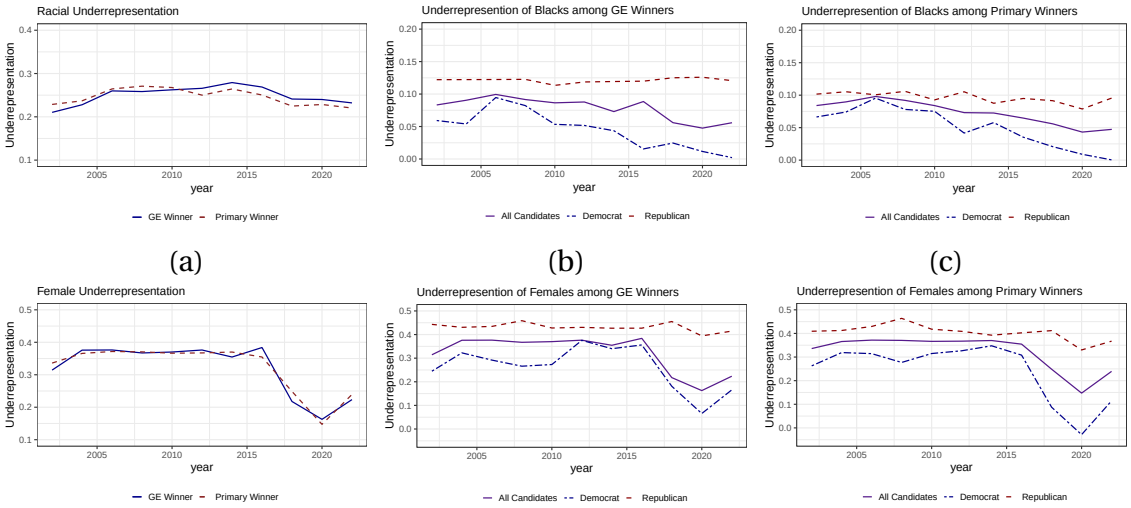
(d) $\delta_q \uparrow$ and $\hat{E}$ (e) $\mu_\xi \uparrow$ and $\hat{E}$ (f) $\sigma_\xi \uparrow$ and $\hat{E}$

Figure 3: This Figure shows how model predictions change when one changes the parameters δ_q, μ_ξ, and σ_ξ. To study this, all potential candidate's expected GE payoff is assumed to be 1 so that we may analyze the changes in the outcomes keeping future payoffs constant. Note that the GE payoff is estimated separately in the GE state. The baseline parameter values are $\mu_{p,P} = 0$ and $\sigma_{p,P} = 1$ for $P \in \{R, D\}$, $\delta_I = 0.5$, $\delta_{q,P} = 0$ for $P \in \{R, D\}$, $\mu_\xi = 0$, and $\sigma_\xi = 0$. Figure 3a shows how the share of primary winners who possess the characteristic $Q^W_{disc,P} = 1$ changes as the weight, $\delta_{q,P}$, associated with that characteristic increases for party P in the large Tullock contest function. Figure 3d shows how the equilibrium mass of entrants changes as $\delta_{q,P}$ is increased. Figure 3b shows how the mean unobserved valence of primary winners from party P changes as μ_ξ varies. Figure 3e shows how the equilibrium mass of entrants from party P changes as μ_ξ varies. Figure 3c shows how the standard deviation/dispersion of unobserved valence changes as σ_ξ varies. Figure 3f shows how the equilibrium mass of entrants from party P changes as σ_ξ varies.



(a) (b)

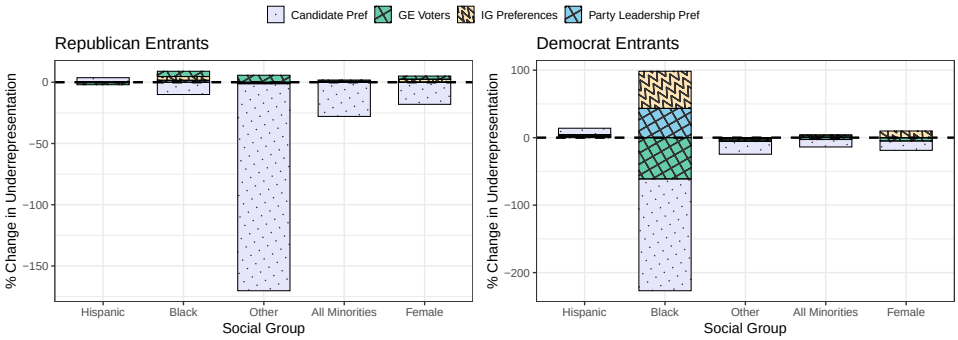
Figure 4: Underrepresentation of Minorities: This figure compares the share of African Americans (Blacks), Hispanics, Other Races, and Females in the general population to their share among primary winners and general election winners. Panel (a) shows the comparison for Primary Winners and panel (b) shows the comparison for general election winners. Gray colored bars refer to share in the population, purple colored bars refer to the shares for all candidates irrespective of party affiliation, blue colored bars refer to shares among Democrats, and red colored bars refer to shares among the Republicans. Note, that representation among the primary winners and general election winners are quite close to one another.



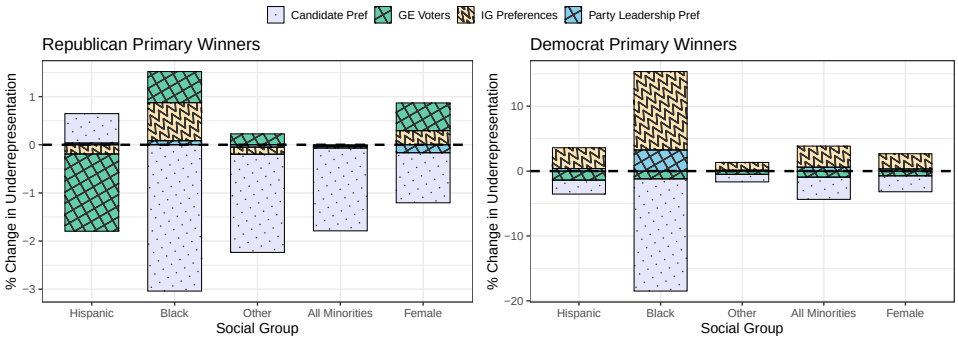
(a) (b) (c)

(d) (e) (f)

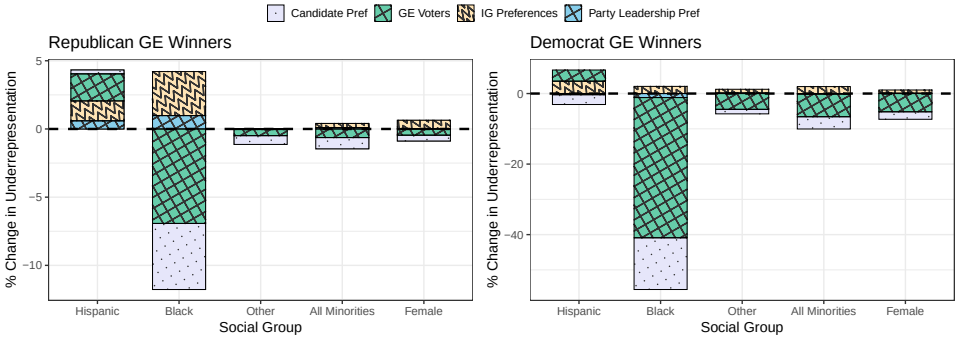
Figure 5: Underrepresentation of Blacks and Females over time: This figure shows the time trend of underrepresentation for African Americans (Blacks) and Females from 2002 to 2022. Here, underrepresentation is defined as $u_{j,t} = p_{j,t} - s_{j,t}$, where $p_{j,t}$ represents the share of social group j in the population in year t, and $s_{j,t}$ represents the share of primary or general election winners from social group j in year t. This quantity reflects the proportion of social group j that lacks representation in primary or general elections. Panel (a) compares the time trend of underrepresentation for all racial minorities between primary winners and general election winners. The red dashed line represents primary winners, while the blue solid line represents general election winners. Panels (b) and (c) plot the underrepresentation trends for Blacks in primary winners and general election winners, respectively. The red dashed line represents the Republican Party, the purple solid line represents all candidates, and the blue line represents the Democratic Party. Panel (d) shows the underrepresentation trend for Females, with the same legend as in panel (a). Panels (e) and (f) present the time trends in Female underrepresentation for primary winners and general election winners, respectively, with the legend matching that of panels (b) and (c).



(a) Entrants

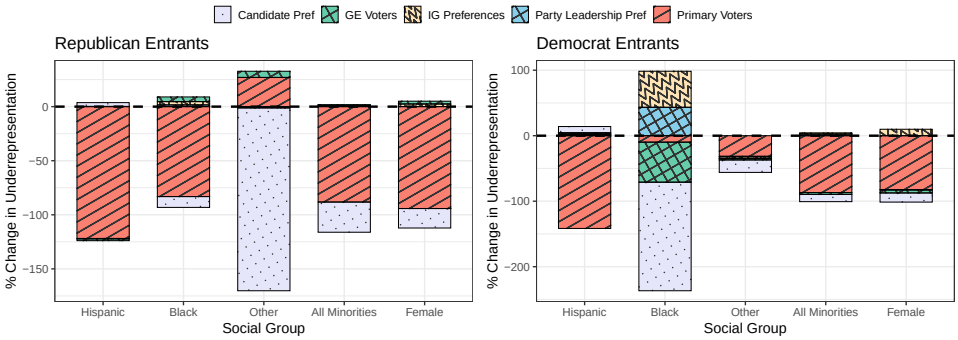


(b) Primary Winner

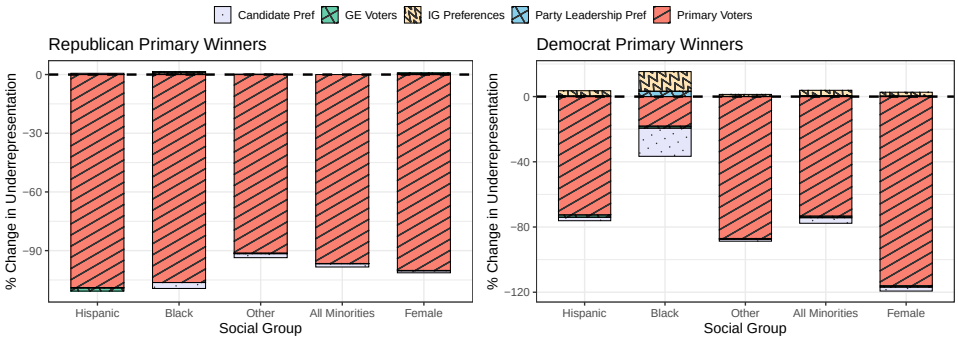


(c) GE Winner

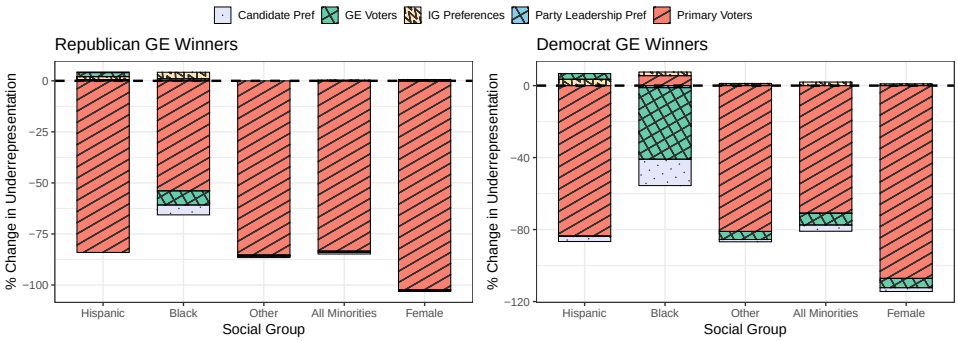
Figure 6: This figure shows the changes in underrepresentation when players and contest functions are made indifferent to the race and gender of candidates. In this figure, the case of primary voters being indifferent is omitted to focus on other factors. Moreover, "GE Voters" is synonymous with "GE contest functions." The y-axis plots proportional change in underrepresentation, defined as $\Delta\text{Underr}_k^{j,l} = \frac{(Prop_k - Q_k^{j,l}) - (Propk - Q_k^{j,o})}{(Prop_k - Q_k^{j,o})}$, where $j \in \{E, PW, GEW\}$ and $k \in$ {Hispanics, Blacks, Other, All Minority races, Female}. $Prop_k$ is the share of social group k in the U.S. population, and $Q_k^{j,l}$ is the predicted share of social group k at stage j (entrants, primary winners, or general election winners) when l is indifferent across race and gender. Here, l indexes the cases when either candidates, interest groups, party leadership, or general election voters are indifferent. Panel (a) shows proportional changes for entrants ($j = E$). Panel (b) for primary winners ($j = PW$). Panel (c) for general election winners ($j = GEW$).



(a) Entrants

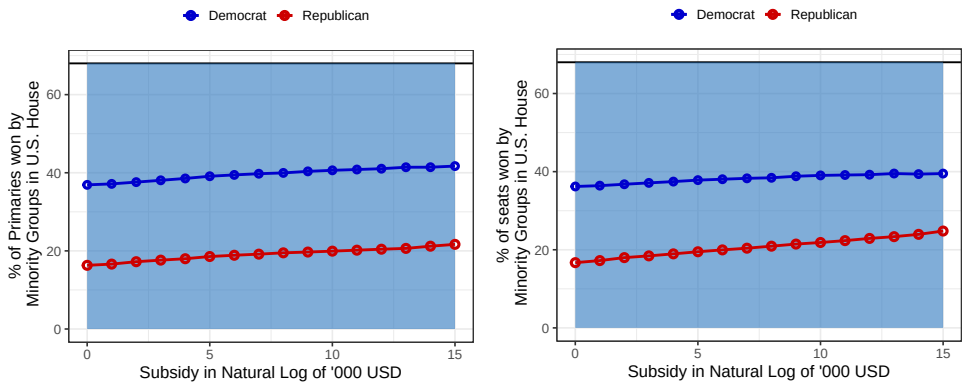


(b) Primary Winner



(c) GE Winner

Figure 7: This figure shows the changes in underrepresentation when players and contest functions are made indifferent to the race and gender of candidates. Moreover, I refer to "GE contest functions" as "GE Voters" and "Primary contest functions" as "Primary Voters". The y-axis plots proportional change in underrepresentation, defined as $\Delta\text{Underr}_k^{j,l} = \frac{(Prop_k - Q_k^{j,l}) - (Propk - Q_k^{j,o})}{(Prop_k - Q_k^{j,o})}$, where $j \in \{E, PW, GEW\}$ and $k \in \{\text{Hispanics, Blacks, Other, All Minority races, Female}\}$. $Prop_k$ is the share of social group k in the U.S. population, and $Q_k^{j,l}$ is the predicted share of social group k at stage j (entrants, primary winners, or general election winners) when l is indifferent across race and gender. Here, l indexes the cases when either candidates, interest groups, party leadership, general election voters, or primary voters are indifferent. Panel (a) shows proportional changes for entrants ($j = E$). Panel (b) for primary winners ($j = PW$). Panel (c) for general election winners ($j = GEW$).



(a) (b)

Figure 8: This figure illustrates how representation, polarization, and average valence changes when campaign support subsidy is provided at the general election stage. Representation is defined as the share of primary or general election winners from underrepresented race-gender pairs, while polarization is measured as the absolute distance between party platform means. The x-axis shows the share of reserved seats in the U.S. House, and in panels (a) and (b), the shaded blue region highlights the area where underrepresentation exists. Panel (a) plots representation measure for Primary Winners by party against subsidy amount of natural log of 1000 USD. Panel (b) plots representation measure for GE Winners by party against subsidy amount of natural log of 1000 USD. In these two figures, the blue region highlights the area where minorities are underrepresented.

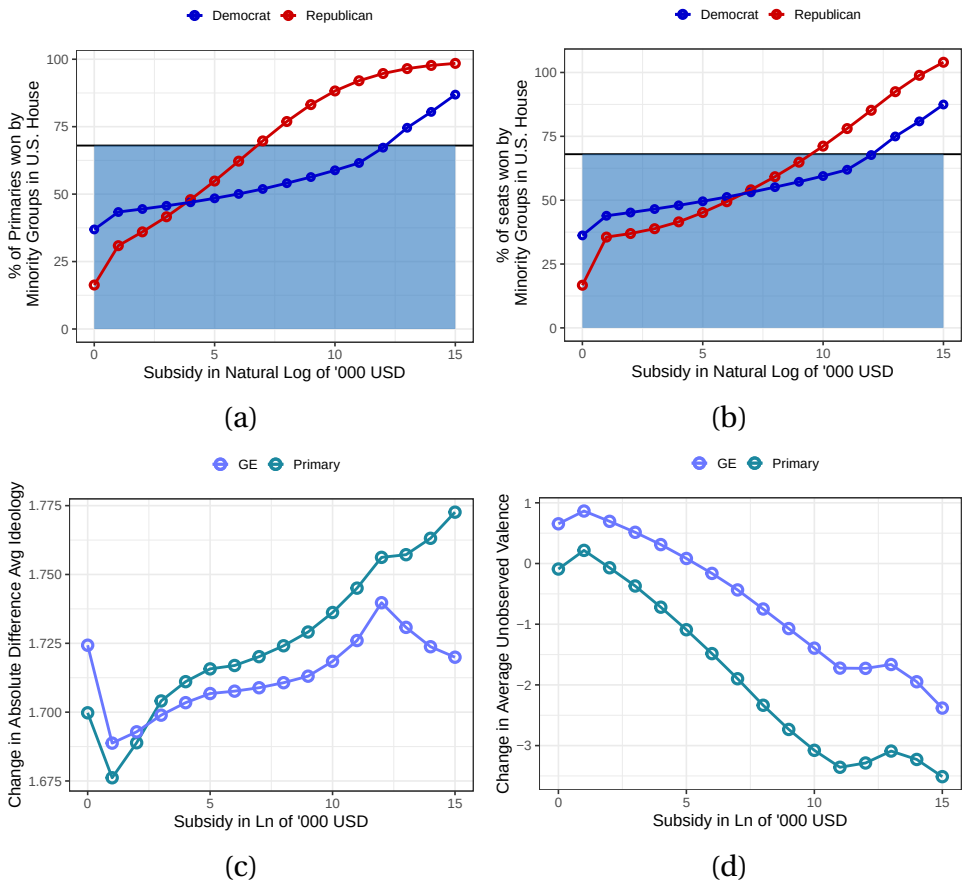


Figure 9: This figure illustrates how representation, polarization, and average valence changes when campaign support subsidy is provided at the general election stage. Representation is defined as the share of primary or general election winners from underrepresented race-gender pairs, while polarization is measured as the absolute distance between party platform means. The x-axis shows the share of reserved seats in the U.S. House, and in panels (a) and (b), the shaded blue region highlights the area where underrepresentation exists. Panel (a) plots representation measure for Primary Winners by party against subsidy amount of natural log of 1000 USD. Panel (b) plots representation measure for GE Winners by party against subsidy amount of natural log of 1000 USD. In these two figures, the blue region highlights the area where minorities are underrepresented. Panel (c) plots changes in absolute difference in party mean ideology among Primary and GE winners against subsidy amount of natural log of 1000 USD. Panel (d) plots changes in average valence of Primary and GE winners against subsidy amount of natural log of 1000 USD.

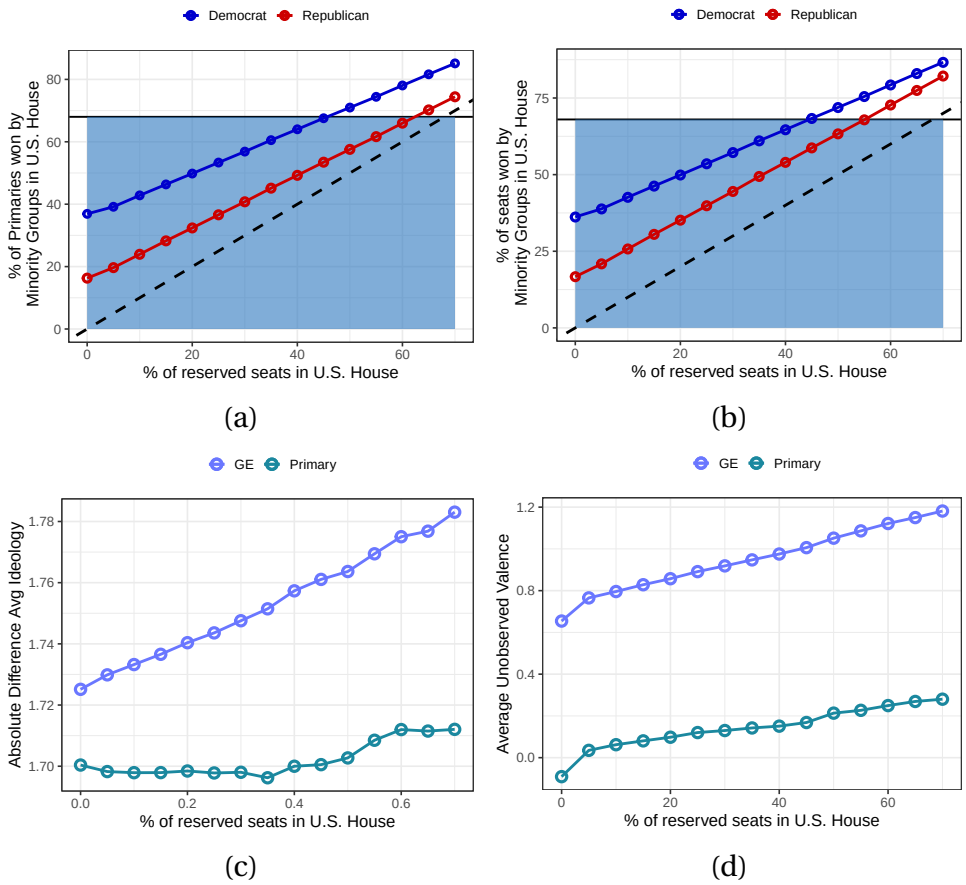


Figure 10: This figure illustrates representation, polarization, and average valence under a quota/reservation policy that allocates a share of seats to underrepresented social groups. Representation is defined as the share of primary or general election winners from underrepresented race-gender pairs, while polarization is measured as the absolute distance between party platform means. The x-axis shows the share of reserved seats in the U.S. House, and the dashed black line represents the 45-degree line. In panels (a) and (b), the shaded blue region highlights the area where underrepresentation exists. Panel (a) plots the representation measure for primary winners by party against the proportion of reserved congressional districts, while Panel (b) does the same for GE winners. Panel (c) plots the change in the absolute difference in party mean ideology among primary and GE winners, and Panel (d) plots changes in the average valence of primary and GE winners, both against the proportion of reserved districts.

10 Tables

Table 1: Summary Statistics

Variable	Mean	Std. Dev.	Max.	Min.	Obs.
Candidate					
White	0.877	0.328	1	0	8288
African American	0.0514	0.221	1	0	8288
Hispanic	0.0434	0.204	1	0	8288
Other Races	0.0279	0.165	1	0	8288
Female	0.188	0.391	1	0	8288
Dime Score	0.149	1.03	2.61	-1.89	8288
GE Campaign Participation					
Candidate Committee	0.77	0.421	1	0	8288
Interest Group Committees	0.439	0.496	1	0	8288
Party Committees	0.115	0.32	1	0	8288
GE Campaign Spending (in 2022 Dollars)					
Candidates	943513.89	1425940.6	24507861	0	8288
Interest Groups	156130.76	926729.07	14569597	0	8288
Party Leadership	102658.47	603927.59	13341610	0	8288
Primary Race					
Number of Contestants	1.97	1.61	19	1	7294
Open	0.411	0.492	1	0	7294
Closed	0.236	0.425	1	0	7294
Semi-Closed	0.236	0.425	1	0	7294
Top-Two	0.117	0.321	1	0	7294
Congressional District Demographics					
Proportion of Male	0.492	0.00671	0.524	0.471	8288
Median Age	37.5	2.74	51.4	28.5	8288
Median Household Income	58700	15200	131000	31500	8288
Proportion Unemployed	0.0349	0.00934	0.0775	0.0152	8288
Proportion without College Degree	0.555	0.058	0.69	0.303	8288
Proportion of Whites	0.742	0.133	0.968	0.202	8288
Proportion of African Americans	0.113	0.0948	0.614	0.00413	8288

Candidates refer to primary winners between 2002-2022 and the summary statistics are provided for those years. *GE Campaign Participation* refers to the extensive margin variation of spending decisions. *GE Campaign Spending* refers to overall (exentsive+intensive) variation in spending levels and these values are deflated to 2022 USD. *Primary Race* data consists for the years 2002-2020, therefore there is a drop in observations. *Congressional District Demographics* are obtained from US Census and ACS.

Table 2: Candidate Characteristics and Demand for Candidates

Dependent Variables:	Party Spending	Interest Group Spending	Vote Shares	Party Spending	Interest Group Spending	Vote Shares
Model:	(1)	(2)	(3)	(4)	(5)	(6)
Variables						
Racial Minority	3.688	-17.00	-0.0083			
	(21.35)	(32.41)	(0.0060)			
African American				-70.87**	-105.0**	-0.0135
				(31.05)	(47.15)	(0.0088)
Hispanics				87.76***	85.69*	-0.0140
				(33.73)	(51.21)	(0.0095)
Other				13.64	-10.69	0.0108
				(41.24)	(62.63)	(0.0117)
Female	24.31	56.26**	-0.0043	25.61	57.81**	-0.0042
	(18.15)	(27.55)	(0.0051)	(18.14)	(27.55)	(0.0051)
Controls	Yes	Yes	Yes	Yes	Yes	Yes
Fixed-effects						
State	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics						
Observations	8,288	8,288	8,288	8,288	8,288	8,288
R^2	0.03125	0.05181	0.02605	0.03277	0.05274	0.02649
Within R^2	0.00439	0.00782	0.02605	0.00595	0.00879	0.02649

*IID standard-errors in parentheses. Level of Significance Codes: ***: 0.01, **: 0.05, *: 0.1.* Regression results for Equation 3.1 on the sample of primary winners between 2002-2022 in U.S. House Congress. Column (1) refers to the regression of party spending on the Minority Dummy. Column (2) refers to the regression of interest group spending on the Minority Dummy. Column (3) refers to the regression of vote shares on the Minority Dummy. Columns (4), (5), and (6) expand the Minority Dummy to individual race categories for columns (1), (2), and (3), respectively. The race categories include African Americans, Hispanics, Others, with White as the left-out category. Each column controls for the following congressional district characteristics: proportion of males, median age, median household income, proportion unemployed, proportion of the adult population (over 25 years old) without a college degree, proportion of whites, and proportion of African Americans.

Table 3: Candidate Characteristics, Party Affiliation, and Platform Positions

Dependent Variables:	Dime Score						Democrat	
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Variables								
Racial Minority	-0.2880***	-0.0697***	0.0382***				0.1410***	
	(0.0385)	(0.0179)	(0.0135)				(0.0185)	
African American				-0.4181***	-0.0066	0.0040		0.2050***
				(0.0560)	(0.0293)	(0.0185)		(0.0270)
Hispanic				-0.1165*	-0.0514**	0.0897***		0.0686**
				(0.0612)	(0.0256)	(0.0228)		(0.0295)
Other				-0.3056***	-0.1887***	0.0435*		0.1319***
				(0.0731)	(0.0338)	(0.0253)		(0.0352)
Female	-0.6356***	0.0375**	-0.0810***	-0.6312***	0.0353**	-0.0795***	0.2880***	0.2859***
	(0.0317)	(0.0163)	(0.0111)	(0.0317)	(0.0162)	(0.0111)	(0.0152)	(0.0152)
Fixed-effects								
State	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Fit statistics								
Observations	7,322	3,688	3,634	7,322	3,688	3,634	7,322	7,322
R^2	0.08217	0.30858	0.23650	0.08398	0.31214	0.23846	0.06497	0.06660
Within R^2	0.07360	0.02608	0.10841	0.07543	0.03109	0.11070	0.06384	0.06547

*IID standard-errors in parentheses. Level of Significance Codes: ***: 0.01, **: 0.05, *: 0.1.* Regression results for Equation 3.2 on the sample of primary winners (and by primary winners by party affiliation) between 2002-2022 in U.S. House Congress. In columns (1) to (6), the dependent variable is the dynamic Dime score obtained from Bonica (2019). Column (1) reports the results using the Minority Dummy for the full sample of primary winners. Column (2) presents the results using the Minority Dummy for the sample of Republican primary winners, while column (3) shows the results for the sample of Democratic primary winners. Columns (4), (5), and (6) expand the Minority Dummy to individual race categories for the full sample, Republican primary winners, and Democratic primary winners, respectively. Column (7) regresses whether a primary winner is affiliated with the Democratic Party on the Minority Dummy for the full sample, while column (8) regresses party affiliation on individual race categories for the full sample. The race categories are African Americans, Hispanics, Others, and White (as the reference category). Each column controls for the following congressional district characteristics: proportion of males, median age, median household income, proportion of unemployed, proportion of adults (aged >25) without a college degree, proportion of whites, and proportion of African Americans.

Table 4: General Election Stage Estimates

Parameter	Estimate	Std. Error	Parameter	Estimate	Std. Error
β	0.0477***	0.00906	$\beta_{q,IG,Hispanic}$	2.68***	0.82
I_{IG}	0.293	0.205	$\beta_{q,IG,White}$	1.94***	0.532
$w_{I,PL}$	1.26***	0.236	$\beta_{q,IG,Black}$	4.3***	0.647
σ_C	7.1***	0.193	$\beta_{q,IG,Others}$	0.606	0.914
σ_{IG}	3.76***	0.102	$\beta_{q,IG,Male}$	-3.56***	0.335
σ_{PL}	2.29***	0.0491	*Party Preferences*		
σ_ξ	2.85***	0.0829	$\beta_{q,PL,Party}$	0.16	0.158
Candidate Preferences			$\beta_{q,PL,Hispanic}$	-4.13***	0.487
$\beta_{q,C,Dem}$	1.65***	0.268	$\beta_{q,PL,White}$	-5.34***	0.446
$\beta_{q,C,Hispanic}$	-4.41***	0.51	$\beta_{q,PL,Black}$	-3.63***	0.449
$\beta_{q,C,White}$	-5.17***	0.541	$\beta_{q,PL,Others}$	-5.76***	0.505
$\beta_{q,C,Black}$	-7.08***	0.582	$\beta_{q,PL,Male}$	-0.153	0.171
$\beta_{q,C,Others}$	-11.8***	1.34	*Voter Preferences*		
$\beta_{q,C,Male}$	1.44***	0.18	$\beta_{q,V,Hispanic}$	0.71**	0.291
IG Preferences			$\beta_{q,V,Black}$	-1.66***	0.265
$\beta_{q,IG,Party}$	-0.38	0.857	$\beta_{q,V,Other}$	-2.13***	0.31
			$\beta_{q,V,Male}$	1.58***	0.168
Observations	4,144		Median Voter Ideology Controls Included, $X_d'\beta_{ideo,v}$		
Penalized Likelihood	34846.64		Conditional Valence Controls Included, $X_d'\beta_\xi$		

*Level of Significance Codes: ***: 0.01, **: 0.05, *: 0.1.* Penalization parameter, λ, is set to 6000. The Monte-Carlo performance of the estimator at this penalization parameter value is reported in Table A3. The units of parameters governing candidate preferences, interest group preferences, and party leadership preferences are natural log of 1,000 USD (deflated to 2022 dollars). Controls for the following congressional district characteristics are present: year, proportion of males, median age, median household income, proportion of unemployed, proportion of adults (aged >25) without a college degree, proportion of whites, and proportion of African Americans. The estimates for the coefficients of these controls are reported in Table A2.

Table 5: General Election Model Fit

	Data	Model		Data	Model
Rep Cand Spending	971.9	972.88	Rep Party Spending	114.34	114.25
Dem Cand Spending	901.56	902.23	Dem Party Spending	89.496	89.508
White Cand Spending	963.53	964.45	Party Spending on White	101.52	101.47
Black Cand Spending	522.72	522.55	Party Spending on Black	27.905	27.899
Male Cand Spending	930.38	931.57	Party Spending on Male	93.65	93.573
Rep IG Spending	172.01	152.87	Rep White Voteshare	0.49826	0.4908
Dem IG Spending	138.01	124.98	Rep Black Voteshare	0.54697	0.45488
IG Spending on White	153.16	127.8	Rep Male Voteshare	0.50593	0.50918
IG Spending on Black	82.93	159.74	Dem White Voteshare	0.50273	0.5163
IG Spending on Male	131.1	110.1	Dem Black Voteshare	0.32045	0.33987
			Dem Male Voteshare	0.49838	0.51902

This table reports the model fit for the general election stage. The spending levels are in the units of 1,000 USD (deflated to 2022 dollars). Note that the observed spending levels are quite closely matched by the model predictions. Moreover, the model does not provide a prediction for vote shares, but the probability of winning an election match closely with the observed vote shares.

Table 6: Entry and Primary Stage Estimates

Parameter	Estimate	Std. Error	Parameter	Estimate	Std. Error
Candidate Dist. Estimates			*Primary Voter Ideal Point Est.*		
μ_{ξ}	-5.51***	1.06	$\delta_{I,year}$	0.27***	0.0861
$\mu_{p,R}$	1.15***	0.0613	$\delta_{I,\text{Prop. of males}}$	0.571***	0.0759
$\mu_{p,D}$	-0.865***	0.0721	$\delta_{I,\text{median age}}$	-0.108	0.0872
σ_{ξ}	3.41***	0.0273	$\delta_{I,\text{median HH income}}$	-0.453***	0.152
$\sigma_{p,R}$	0.451***	0.036	$\delta_{I,\text{Prop unemployed}}$	0.0537	0.0623
$\sigma_{p,D}$	0.123***	0.0356	$\delta_{I,\text{less than college}}$	-0.48***	0.151
Primary Voter Pref. over Q			$\delta_{I,\text{Prop of white}}$	0.664***	0.204
$\delta_{q,R,Hispanic}$	-4.71***	0.4	$\delta_{I,\text{Prop of black}}$	0.494***	0.128
$\delta_{q,R,White}$	-0.758	0.848	$\delta_{I,D,\text{ lag presid share}}$	-0.0277	0.128
$\delta_{q,R,Black}$	-3.8***	0.275	$\delta_{I,R,\text{ lag presid share}}$	-0.461***	0.0954
$\delta_{q,R,Male}$	3.21***	0.338	$\delta_{I,D,\text{ Open}}$	1.51***	0.098
$\delta_{q,D,Hispanic}$	-0.307	1.16	$\delta_{I,R,\text{ Open}}$	0.738***	0.0898
$\delta_{q,D,White}$	3.44***	0.325	$\delta_{I,D,\text{ Closed}}$	-2.7***	0.24
$\delta_{q,D,Black}$	2.09***	0.735	$\delta_{I,R,\text{ Closed}}$	-0.988***	0.263
$\delta_{q,D,Male}$	2.27*	1.32	$\delta_{I,D,\text{ Semi-closed}}$	-3.38***	0.201
			$\delta_{I,R,\text{Semi-closed}}$	-0.056	0.294
Observations	3,647				
Objective Function	16.504				

*Level of Significance Codes: ***: 0.01, **: 0.05, *: 0.1.* The Monte-Carlo performance of the estimator used to estimate this stage is reported in Table A4. Median voter ideology, $I_d^{Prim} = X'_{d,Prim}\delta_I$, includes the following covariates: year, proportion of males, median age, median household income, proportion of unemployed, proportion of adults (aged >25) without a college degree, proportion of whites, proportion of African Americans, lagged presidential vote share × Party, open primary × Party, closed primary × Party , and semi-closed primary× Party.

Table 7: Model Fit

	Data	Model		Data	Model
Hispanic Winners R	0.0362	0.0498	Black Winners D	0.0691	0.0568
White Winners R	0.917	0.903	Other Winners D	0.0345	0.0282
Black Winners R	0.028	0.0262	Male Winners D	0.744	0.734
Other Winners R	0.0192	0.0211	Mean Ideo R	1.00	1.00
Male Winners R	0.894	0.875	Mean Ideo D	-0.7	-0.735
Hispanic Winners D	0.0474	0.0749	Average # Rep Entrants	2.09	1.94
White Winners D	0.849	0.84	Average # Dem Entrants	1.85	1.65

This table reports the model fit for the Entry and Primary stage, assessing whether the model's predicted shares of primary winners across party affiliation and social groups align with their observed counterparts.

References

Acemoglu, D., G. Egorov, and K. Sonin (2010). Political selection and persistence of bad governments. *The Quarterly Journal of Economics 125*(4), 1511–1575.

Acharya, A., E. Grillo, T. Sugaya, and E. Turkel (2022). Electoral campaigns as dynamic contests.

Allen, W. R. and R. Farley (1986). The shifting social and economic tides of black america, 1950-1980. *Annual review of sociology*, 277–306.

Ang, D. (2019). Do 40-year-old facts still matter? long-run effects of federal oversight under the voting rights act. *American Economic Journal: Applied Economics 11*(3), 1–53.

Anzia, S. F. and C. R. Berry (2011). The jackie (and jill) robinson effect: Why do congresswomen outperform congressmen? *American Journal of Political Science 55*(3), 478–493.

Ashworth, S., C. R. Berry, and E. Bueno de Mesquita (2024). Modeling theories of women's underrepresentation in elections. *American Journal of Political Science 68*(1), 289–303.

Avis, E., C. Ferraz, F. Finan, and C. Varjão (2022). Money and politics: The effects of campaign spending limits on political entry and competition. *American Economic Journal: Applied Economics 14*(4), 167–199.

Bajari, P., H. Hong, and S. P. Ryan (2010). Identification and estimation of a discrete game of complete information. *Econometrica 78*(5), 1529–1568.

Beach, B., D. B. Jones, T. Twinam, and R. Walsh (2018). Minority representation in local government. Technical report, National Bureau of Economic Research.

Bernini, A., G. Facchini, M. Tabellini, and C. Testa (2023). Black empowerment and white mobilization: the effects of the voting rights act. Technical report, National Bureau of Economic Research.

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

Besley, T., T. Persson, and D. M. Sturm (2010). Political competition, policy and growth: theory and evidence from the us. *The Review of Economic Studies 77*(4), 1329–1352.

Billings, S. B., N. Braun, D. B. Jones, and Y. Shi (2024). Disparate racial impacts of shelby county v. holder on voter turnout. *Journal of Public Economics 230*, 105047.

Bombardini, M. and F. Trebbi (2011). Votes or money? theory and evidence from the us congress. *Journal of Public Economics 95*(7-8), 587–611.

Bonica, A. (2019). Database on ideology, money in politics, and elections: pre-release version 3.0 [computer file].

Burrell, B. (2014). *Gender in Campaigns for the US House of Representatives*. University of Michigan Press.

Canen, N., C. Kendall, and F. Trebbi (2020). Unbundling polarization. *Econometrica 88*(3), 1197–1233.

Canen, N. J., C. Kendall, and F. Trebbi (2021). Political parties as drivers of us polarization: 1927-2018. Technical report, National Bureau of Economic Research.

Casas-Zamora, K. (2005). *Paying for democracy: political finance and state funding for parties*. ECPR Press.

Cascio, E. U. and E. Washington (2014). Valuing the vote: The redistribution of voting rights and state funds following the voting rights act of 1965. *The Quarterly Journal of Economics 129*(1), 379–433.

Clayton, A. (2021). How do electoral gender quotas affect policy? *Annual Review of Political Science 24*(1), 235–252.

Cox, C. (2022). Campaign finance in the age of super pacs. *Available at SSRN 3794817*.

Dal Bó, E. and F. Finan (2018). Progress and perspectives in the study of political selection. *Annual Review of Economics 10*(1), 541–575.

Dal Bó, E., F. Finan, O. Folke, T. Persson, and J. Rickne (2017). Who becomes a politician? *The Quarterly Journal of Economics 132*(4), 1877–1914.

Davidson, C. and G. Korbel (1981). At-large elections and minority-group representation: A re-examination of historical and contemporary evidence. *The Journal of Politics 43*(4), 982–1005.

Desai, Z., V. Karekurve-Ramachandra, and S. Montero (2024). Are women better politicians? discrimination, gender quotas, and electoral accountability.

Diermeier, D., M. Keane, and A. Merlo (2005). A political economy model of congressional careers. *American Economic Review 95*(1), 347–373.

Fox, R. L. and J. L. Lawless (2004). Entering the arena? gender and the decision to run for office. *American journal of political science 48*(2), 264–280.

Fox, R. L. and J. L. Lawless (2011). Gendered perceptions and political candidacies: A central barrier to women's equality in electoral politics. *American journal of political science 55*(1), 59–73.

Frey, W. H. (2018). *Diversity explosion: How new racial demographics are remaking America*. Brookings Institution Press.

Hassell, H. J. and N. Visalvanich (2019). The party's primary preferences: Race, gender, and party support of congressional primary candidates. *American Journal of Political Science 63*(4), 905–919.

Hirano, S., J. M. Snyder Jr, et al. (2014). Primary elections and the quality of elected officials. *Quarterly Journal of Political Science 9*(4), 473–500.

Huang, Y. and M. He (2021). Structural analysis of tullock contests with an application to us house of representatives elections. *International Economic Review 62*(3), 1011–1054.

Iaryczower, M., G. Lopez-Moctezuma, and A. Meirowitz (2022). Career concerns and the dynamics of electoral accountability. *American Journal of Political Science*.

Iaryczower, M., S. Montero, and G. Kim (2022). Representation failure. Technical report, National Bureau of Economic Research.

Jha, A. (2023). Rally the vote: Electoral competition with direct campaign communication. *Available at SSRN 4598283*.

Johnson, K. M. and D. T. Lichter (2016). Diverging demography: Hispanic and non-hispanic contributions to us population redistribution and diversity. *Population Research and Policy Review 35*, 705–725.

Kang, K. (2016). Policy influence and private returns from lobbying in the energy sector. *The Review of Economic Studies 83*(1), 269–305.

Kanthak, K. and J. Woon (2015). Women don't run? election aversion and candidate entry. *American journal of political science 59*(3), 595–612.

Kawai, K. and T. Sunada (2022). Estimating candidate valence. Technical report, National Bureau of Economic Research.

Lawless, J. L. and R. L. Fox (2010). *It still takes a candidate: Why women don't run for office*. Cambridge University Press.

Lawless, J. L. and K. Pearson (2008). The primary reason for women's underrepresentation? reevaluating the conventional wisdom. *The Journal of Politics 70*(1), 67–82.

Marschall, M. J., A. V. Ruhil, and P. R. Shah (2010). The new racial calculus: Electoral institutions and black representation in local legislatures. *American Journal of Political Science 54*(1), 107–124.

Matsusaka, J. G. and C. Kendall (2021). The common good and voter polarization. *USC CLASS Research Paper No. CLASS21-38, USC Law Legal Studies Paper* (21-38).

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica: Journal of the Econometric Society*, 995–1026.

Newey, W. K. and D. McFadden (1994). Large sample estimation and hypothesis testing. *Handbook of econometrics 4*, 2111–2245.

Poole, K. T. and H. Rosenthal (1985). A spatial model for legislative roll call analysis. *American journal of political science*, 357–384.

Ricca, F. and F. Trebbi (2022). *Minority underrepresentation in US cities*. Number w29738. National Bureau of Economic Research.

Rosen, J. (2017). Gender quotas for women in national politics: A comparative analysis across development thresholds. *Social Science Research 66*, 82–101.

Schuit, S. and J. C. Rogowski (2017). Race, representation, and the voting rights act. *American Journal of Political Science 61*(3), 513–526.

Shah, P. (2014). It takes a black candidate: A supply-side theory of minority representation. *Political Research Quarterly 67*(2), 266–279.

Shah, P. R., M. J. Marschall, and A. V. Ruhil (2013). Are we there yet? the voting rights act and black representation on city councils, 1981–2006. *The Journal of Politics 75*(4), 993–1008.

Shelby County v. Holder (2013). Shelby county v. holder. Decided June 25, 2013.

Strömberg, D. (2008). How the electoral college influences campaigns and policy: The probability of being florida: American economic review, 98(3), 769-807. *American Economic Review 98*, 769–807.

Tamer, E. (2003). Incomplete simultaneous discrete response model with multiple equilibria. *The Review of Economic Studies 70*(1), 147–165.

Trebbi, F., P. Aghion, and A. Alesina (2008). Electoral rules and minority representation in us cities. *The Quarterly Journal of Economics 123*(1), 325–357.

Trounstine, J. (2010). Representation and accountability in cities. *Annual Review of Political Science 13*(1), 407–423.

Trounstine, J. and M. E. Valdini (2008). The context matters: The effects of single-member versus at-large districts on city council diversity. *American Journal of Political Science 52*(3), 554–569.

Warshaw, C. (2019). Local elections and representation in the united states. *Annual Review of Political Science 22*(1), 461–479.

A Proofs

A.1 Lemma A.1

Lemma A.1 *Given Assumption 2.3, the following holds for* $P \in \{R, D\}$, $X_{d,Prim} \in \mathbb{R}^{K_{X,Prim}}$, $\delta_q \in \mathbb{R}^{K_Q}$, *and* $\delta_I \in \mathbb{R}^{K_X}$

$$\bar{A} = \int_{Q,p,\xi} e^{Q'\delta_q - \left(p - X'_{d,Prim}\delta_I\right)^2 + \xi} dF_P(Q,p,\xi) < \infty \tag{A.1}$$

Proof: Note using independence of $Q_{cont}, Q_{disc}, p, \xi$ one can re-write the integral as:

$$\begin{aligned}
&\int_{Q,p,\xi} e^{Q'\delta_q - \left(p - X'_{d,Prim}\delta_I\right)^2 + \xi} dF_P(Q,p,\xi) = \\
&\int_{Q_{cont}} e^{Q'_{cont}\delta_{q,cont}} dF(Q_{cont}) \times \prod_{l=1}^{K_{disc}} \left(\sum_{x \in \mathcal{Q}_{l,disc}} e^{x_l \delta_{l,disc}} q_{x,l,disc} \right) \\
&\times \int_p e^{-\left(p - X'_{d,Prim}\delta_I\right)^2} dF_P(p) \times \int_{Q_{cont}} e^{\xi} dF_P(\xi).
\end{aligned} \tag{A.2}$$

First consider $\int_{Q_{cont}} e^{Q'_{cont}\delta_{q,cont}} dF(Q_{cont}) = \frac{1}{\sigma_z} \int_z e^z \cdot \phi\left((z - \mu_z)/\sigma_z\right) dz = e^{\mu_z + \frac{\sigma_z^2}{2}} < \infty$, where $\mu_z = \delta'_{cont}\mu_{cont}$, $\sigma_z = \sqrt{\delta'_{cont}\Sigma_{cont}\delta_{cont}}$, and $\phi()$ is the standard normal pdf. Similarly, $\int_{Q_{cont}} e^{\xi} dF_P(\xi) = e^{\mu_\xi + \frac{\sigma_\xi^2}{2}} < \infty$. Moreover, $\sum_{x \in \mathcal{Q}_{l,disc}} e^{x_l \delta_{l,disc}} q_{x,l,disc} < \infty$. Now to show that $\int_p e^{-\left(p - X'_{d,Prim}\delta_I\right)^2} dF_P(p) < \infty$ consider the following

$$\begin{aligned}
&\int_p \exp\left\{ -\left(p - X'_{d,Prim}\delta_I\right)^2 \right\} dF_P(p) \\
&= \frac{1}{\sqrt{2\pi}\sigma_{p,P}} \int_p \exp\left\{ -\left(p - X'_{d,Prim}\delta_I\right)^2 - \frac{1}{2\sigma_{p,P}^2}\left(p - \mu_{p,P}\right)^2 \right\} dp \\
&= \frac{1}{\sqrt{2\pi}\sigma_{p,P}} \cdot \exp\left\{ \frac{\mu_{p,P}^2}{2\sigma_{p,P}^2} + \left(X'_{d,Prim}\delta_I\right)^2 - \left(\frac{\mu_{p,P}}{2\sigma_{p,P}^2} + \left(X'_{d,Prim}\delta_I\right) \right)^2 \frac{2\sigma_{p,P}^2}{1 + 2\sigma_{p,P}^2} \right\}
\end{aligned} \tag{A.3}$$

$$\times \int_p \exp\left\{ -\frac{1}{2} \left(\frac{p - \left(\frac{\mu_{p,P}}{2\sigma_{p,P}^2} + X'_{d,Prim}\delta_I \right) \cdot \frac{2\sigma_{p,P}}{1+2\sigma_{p,P}^2}}{\sqrt{\frac{2\sigma_{p,P}^2}{2\sigma_{p,P}^2+1}}} \right)^2 \right\} dp \tag{A.4}$$

$$= \sqrt{\frac{2}{2\sigma_{p,P}^2 + 1}} \cdot \exp\left\{ \frac{\mu_{p,P}^2}{2\sigma_{p,P}^2} + \left(X'_{d,Prim}\delta_I\right)^2 - \left(\frac{\mu_{p,P}}{2\sigma_{p,P}^2} + \left(X'_{d,Prim}\delta_I\right) \right)^2 \frac{2\sigma_{p,P}^2}{1 + 2\sigma_{p,P}^2} \right\} < \infty$$

The first equality holds by substituting the probability density function of $F_p(p)$ which is a normal distribution with mean $\mu_{p,P}$ and standard deviation $\sigma_{p,P}$. The second equality is obtained by first completing the squares to obtain a single quadratic in terms of p and then

taking out the terms that do not depend on p out of the integral. The last equality holds because the integral is of the form $\int_z e^{-(\frac{z-\mu_z}{\sigma_z})^2/2} dz$ and therefore it is equal to $\sqrt{2\pi}\sigma_z$, which is then substituted and simplified. In the last term expression each individual term is finite which implies that the whole integral is also finite.

A.2 Proof of Proposition 2.2

First note that the equilibrium entry strategy π^s follows from the Program 2.7. Therefore, it is suffice to prove that there is a unique solution to the equation 2.13. Consider the probability density function,

$$g_P(Q,p,\xi) = \frac{1}{\bar{A}} e^{Q'\delta_q - \left(p - X'_{d,Prim}\delta_I\right)^2 + \xi} \cdot f_P(Q,p\xi). \tag{A.5}$$

Note, given this we can re-write equation 2.2 as followed:

$$\begin{aligned} \bar{A}\int_{Q,p,\xi} \pi^s g_P(Q,p,\xi)\, d(Q,p,\xi) - A = 0, \\ \Rightarrow \bar{A}\sum_{\mathbf{x}\in Q_{disc}} \int_{Q_{cont},p,\xi} \pi^s \cdot g_P(Q_{cont},p,\xi)\, d(Q_{cont},p,\xi)\, g_{disc}(\mathbf{x}) - A = 0. \end{aligned} \tag{A.6}$$

Now define V_x as

$$V_x(Q_{cont},p,\xi) = \exp\left\{\mathbf{x}'\delta_{q,disc} + Q'\delta_{q,cont} - (p - X'_{d,Prim}\delta_I)^2 + \xi\right\} \cdot \mathbb{E}\left[V_C^{II}(\mathbf{W}_d, \boldsymbol{\xi}_d)\middle| Q_i, p_i, \xi_i, E^O_{j,d}, X_d\right]. \tag{A.7}$$

Moreover let $G_{V,P,\mathbf{x}}$ be the cumulative distribution function of V defined as:

$$G_{V,P,\mathbf{x}}(V) = \int_{V_x(Q_{cont},p,\xi)\le V} g_{P,cont}(Q_{cont},p,\xi)\, d(Q_{cont},p,\xi). \tag{A.8}$$

where $g_{P,cont} = g_P/g_{disc}(\mathbf{x})$. Here $g_{disc}(\mathbf{x}) = \prod_l q_{x,l,disc}$. Also note, Q_{cont}, p and ξ are continuous random variables and $\mathbb{E}\left[V_C^{II}(\mathbf{W}_d, \boldsymbol{\xi}_d)\middle| Q_i, p_i, \xi_i, E^O_{j,d}, X_d\right]$ is also a continuously differentiable function. Then, given $\mathbf{x}$, $V_{\mathbf{x}}(Q_{cont},p,\xi)$ is also a continuous random variable. Therefore, $G_{V,R,x}$ is a continuous function.[32] Now we can re-write the integral as:

$$\begin{aligned} \bar{A}\sum_{\mathbf{x}\in Q_{disc}} \int_V \mathbb{1}\left\{\frac{V}{A} - \kappa > 0\right\} dG_{V,P,\mathbf{x}}(V)\cdot g_{disc}(\mathbf{x}) - A = 0, \\ \Rightarrow \bar{A}\sum_{\mathbf{x}\in Q_{disc}} (1 - G_{V,P,\mathbf{x}}(A\kappa))\cdot g_{disc}(\mathbf{x}) - A = 0. \end{aligned} \tag{A.9}$$

Define $\Gamma(A) = \bar{A}\sum_{\mathbf{x}\in Q_{disc}} (1 - G_{V,P,\mathbf{x}}(A\kappa))\cdot g_{disc}(\mathbf{x}) - A$. Note that $\lim_{A\to\infty}\Gamma(A) \to -\infty < 0$ and $\lim_{A\to 0}\Gamma(A) = \bar{A} > 0$. Therefore, there should be at least one solution to the equation by intermediate value theorem. Finally, note $G_{V,P,x}(A\kappa)$ is weakly increasing since it is cumulative distribution function and V is a continuous random variable. Moreover, $-A$ term

[32] It may not be absolutely continuous and fail to have a pdf, however $G_{V,P,x}$ is continuous.

is strictly decreasing making $\Gamma(A) = \bar{A}\sum_{x\in Q_{disc}}(1 - G_{V,P,x}(A\kappa))g_{disc}(x) - A$ to be a strictly decreasing function. This implies that $\Gamma(A)$ must cross 0 only once, meaning that there is a unique solution. Hence, we have a unique equilibrium.

B Identification of GE Stage Parameters

For exposition, I abstract away from observable district characteristics, assume that all districts observe positive level of spending by all players, and focus only on two race categories. The same arguments are applicable to gender. I also ignore the policy preferences of players and voters as these can be incorporated since I observe candidate platform positions. Also note that all throughout the paper I assume that the econometrician knows γ, which I calibrate to $1/2$ (Cox, 2022). The econometrician observes spending levels and Republican win probabilities $\{S_{R,l,d}, S_{D,l,d}, P_{R,d}\}_{l\in\{C,IG,PL\}}$ for $d = 1, 2, \ldots, D$. The unknown parameters are the Republican bias (b_{Rep}), spending effectiveness (β), candidate unobserved valence ($\xi_{i,d}$), value player l associates with a candidate of social identity j winning the office ($\beta_{q,j,l}$), and voter discrimination (for now assume white v. black b_{wb}). Note that election aversion is captured by $\beta_{q,w,C} - \beta_{q,b,C}$, that is the difference b/w the payoffs that white and black candidates receive. I make the following assumptions

Assumption B.1 *I impose the following assumptions on cost shocks and unobserved valence shocks.*

1 *Idiosyncratic costs shocks have mean zero within each districts, d, conditional on valence and race of candidates.*

$$\mathbb{E}_l\left[\xi_{cost,i,l,d}|\xi_{i,d}, Q_{R,d}, Q_{D,d}\right] = 0 \tag{B.1}$$

where the operator $\mathbb{E}_l$ takes expectations over players l.

2 *Expected Unobserved valence of candidates in districts with same race candidates, is constant across parties.*

$$\mathbb{E}\left[\xi_{R,d} - \xi_{D,d}|Q_{R,d} = Q_{D,d}\right] = 0 \tag{B.2}$$

3 *For $d = 1$, where both competing candidates are of the same race $\xi_{R,1} = 0$.*

Now first consider districts of type A and B where candidates share the same identity as white ($Q_{R,d} = Q_{D,d} = w$) and black ($Q_{R,d} = Q_{D,d} = b$) respectively. The first order conditions

for districts of type A and B, $l \in \{C, IG, PL\}$, and $j \in \{w, b\}$:

$$\exp\left(\sum_{j\in\{w,b\}} \beta_{q,j,l} \cdot \mathbb{1}\{Q_{R,d} = j\} + \xi_{R,d} + \xi_{cost,R,l,d}\right) \cdot \gamma S_{R,l,d}^{\gamma-1} \frac{1 - h_d + \beta \sum_l S_{D,l,d}^{\gamma}}{\left(1 + \beta \sum_k S_{R,k,d}^{\gamma} + \beta \sum_k S_{D,k,d}^{\gamma}\right)^2} = 1 \quad \text{(B.3)}$$

$$\exp\left(\sum_{j\in\{w,b\}} \beta_{q,j,l} \cdot \mathbb{1}\{Q_{R,d} = j\} + \xi_{D,d} + \xi_{cost,D,l,d}\right) \cdot \gamma S_{D,l,d}^{\gamma-1} \frac{h_d + \beta \sum_k S_{R,k,d}^{\gamma}}{\left(1 + \beta \sum_k S_{R,k,d}^{\gamma} + \beta \sum_k S_{D,k,d}^{\gamma}\right)^2} = 1, \quad \text{(B.4)}$$

where w stands for *white* and b stands for *black*. Moreover, note that h_d in these districts is given by $h_d = logistic\left(b_{Rep} + \Delta_i \xi_{i,d}\right)$. Here b_{Rep} represents Republcian bias, this term essentially contains the policy preferences of the district, for exposition I assume it is constant. Now consider the ratios of these two equations. This gives us

$$\exp\left(\Delta_i \xi_{i,d} + \Delta_i \xi_{cost,i,l,d}\right) \cdot \left(\frac{S_{R,l,d}}{S_{D,l,d}}\right)^{\gamma-1} \cdot \frac{1 - h_d + \beta \sum_l S_{D,l,d}^{\gamma}}{h_d + \beta \sum_k S_{R,k,d}^{\gamma}} = 1 \quad \text{(B.5)}$$

$$\Rightarrow \log\left(\left(\frac{S_{R,l,d}}{S_{D,l,d}}\right)^{\gamma-1} \cdot \frac{1 - h_d + \beta \sum_l S_{D,l,d}^{\gamma}}{h_d + \beta \sum_k S_{R,k,d}^{\gamma}}\right) = -\Delta_i \xi_{i,d} - \Delta_i \xi_{cost,i,l,d} \quad \text{(B.6)}$$

Also note that the winning probability, observed by the econometrician, satisfies

$$P_R = \frac{h_d + \beta \sum_k S_{R,k,d}^{\gamma}}{1 + \beta \sum_l S_{D,l,d}^{\gamma} + \beta \sum_l S_{R,l,d}^{\gamma}} \quad \text{(B.7)}$$

$$\Rightarrow logistic\left(b_{Rep} + \Delta_i \xi_{i,d}\right) \equiv h_d = P_{R,d} \cdot \left(1 + \beta \sum_l S_{D,l,d}^{\gamma} + \beta \sum_l S_{R,l,d}^{\gamma}\right) - \beta \sum_k S_{R,k,d}^{\gamma} \quad \text{(B.8)}$$

Note that $\xi_{cost,i,l,d}$ are idiosyncratic shocks and therefore satisfy, $\mathbb{E}_l\left[\Delta_i \xi_{cost,i,l,d} | \xi_{i,d}, Q_{R,d}, Q_{D,d}\right] = 0$ for all d. Then substituting h_d in equation B.6 and applying $\mathbb{E}_l\left[.|\xi_{i,d}, Q_{R,d}, Q_{D,d}\right]$ on both sides gives me,

$$\Delta_i \xi_{i,d} = -\mathbb{E}_l\left[\log\left(\left(\frac{S_{R,l,d}}{S_{D,l,d}}\right)^{\gamma-1} \cdot \frac{(1 - P_{R,d}) \cdot \left(1 + \beta \sum_l S_{D,l,d}^{\gamma} + \beta \sum_l S_{R,l,d}^{\gamma}\right)}{P_{R,d} \cdot \left(1 + \beta \sum_l S_{D,l,d}^{\gamma} + \beta \sum_l S_{R,l,d}^{\gamma}\right) - \beta \sum_k S_{R,k,d}^{\gamma} + \beta \sum_k S_{R,k,d}^{\gamma}}\right)\right]. \quad \text{(B.9)}$$

Finally, note that $\mathbb{E}\left[\Delta_i \xi_{i,d} | Q_{R,d} = Q_{D,d}\right] = 0$. This does not imply that white have the same unobserved valence as blacks, but rather the expected difference of valence within competing candidates that share the same social identity is zero. This gives me the following equation to recover β

$$\mathbb{E}\left[\mathbb{E}_l\left[\log\left(\left(\frac{S_{R,l,d}}{S_{D,l,d}}\right)^{\gamma-1} \cdot \frac{(1 - P_{R,d}) \cdot \left(1 + \beta \sum_l S_{D,l,d}^{\gamma} + \beta \sum_l S_{R,l,d}^{\gamma}\right)}{P_{R,d} \cdot \left(1 + \beta \sum_l S_{D,l,d}^{\gamma} + \beta \sum_l S_{R,l,d}^{\gamma}\right) - \beta \sum_k S_{R,k,d}^{\gamma} + \beta \sum_k S_{R,k,d}^{\gamma}}\right)\right] \middle| Q_{R,d} = Q_{D,d}\right] = 0 \quad \text{(B.10)}$$

Then substituting β in equation B.9 recovers $\Delta_i \xi_{i,d}$. Substituting $\Delta_i \xi_{i,d}$ and β in equation B.8 recovers b_{Rep}. Now, once we have these three objects, we can recover $\beta_{q,l}$ by using the FOC for Republican candidate. Note that the FOC can be re-written as

$$-\log\left(\gamma S_{R,l,d}^{\gamma-1} \frac{1 - h_d + \beta \sum_l S_{D,l,d}^{\gamma}}{\left(1 + \beta \sum_k S_{R,k,d}^{\gamma} + \beta \sum_k S_{D,k,d}^{\gamma}\right)^2}\right) \equiv \tilde{Y}_{R,l,d} = \sum_{j\in\{w,b\}} \beta_{q,j,l} \cdot \mathbb{1}\{Q_{R,d} = j\} + \xi_{R,d} + \xi_{cost,R,l,d} \quad \text{(B.11)}$$

Note that the variable $\tilde{Y}_{R,l,d}$ only requires knowledge of parameters β, b_{Rep}, and $\Delta\xi_{i,d}$ which are recovered from equations B.10, B.8, and B.9. Re-writing the equation as,

$$\begin{aligned}\tilde{Y}_{R,l,d} &= \sum_k \sum_{j\in\{w,b\}} \beta_{q,j,k} \cdot \mathbb{1}\{Q_{R,d} = j\} \cdot \mathbb{1}\{k = l\} + \xi_{R,d} + \xi_{cost,R,l,d} \\ &\Rightarrow \tilde{Y}_{R,l,d} = Z'_{l,d} \cdot \beta_q + \xi_{R,d} + \xi_{cost,R,l,d}\end{aligned} \tag{B.12}$$

where $Z_{l,d}$ and β_q are vectors of length $2L$. Here L is the number of players aligned with each side. For ease of notation, $l = C \iff l^o = 1$, $l = IG \iff l^o = 2$, and $l = PL \iff l^o = 3$. Moreover, element of $Z_{l,d}$ at the position $2k-1$ is given as $Z_{l,d,2k-1} = \mathbb{1}\{Q_{R,d} = w\}\cdot\mathbb{1}\{2k-1 = l^o\}$ and the element at position $2k$ is given as $Z_{l,d,2k} = \mathbb{1}\{Q_{R,d} = b\} \cdot \mathbb{1}\{2k = l^o\}$ for $k = 1, 2, \ldots, L$. Note the following holds:

$$\begin{aligned}\tilde{Y}_{R,l,d} - \bar{Y}_{R,d} =& (Z_{l,d} - \bar{Z}_d)' \cdot \beta_q + \xi_{cost,R,l,d} \\ \Rightarrow \beta_q =& \mathbb{E}\left[(Z_{l,d} - \bar{Z}_d)(Z_{l,d} - \bar{Z}_d)' | Q_{R,d} = Q_{D,d}\right]^{-1} \\ & \cdot \mathbb{E}\left[(Z_{l,d} - \bar{Z}_d)(\tilde{Y}_{R,l,d} - \bar{Y}_{R,d}) | Q_{R,d} = Q_{D,d}\right]\end{aligned} \tag{B.13}$$

where $\bar{Y}_{R,d} = \mathbb{E}_l\left[\tilde{Y}_{R,l,d} | Q_{R,d} = Q_{D,d}\right]$ and $\bar{Z}_d = \mathbb{E}_l\left[Z_{l,d} | Q_{R,d} = Q_{D,d}\right]$. Also recall that by assumption $\mathbb{E}_l\left[\xi_{cost,R,l,d} | \xi_{R,d} Q_{R,d} = Q_{D,d}\right] = 0 \Rightarrow \mathbb{E}_l\left[(Z_{l,d} - \bar{Z}_d)\xi_{cost,R,l,d} | Q_{R,d} = Q_{D,d}\right] = 0$. Then we have the degree of election aversion for white v. black, $\beta_{q,w,C} - \beta_{q,b,C} = \beta_{q,1} - \beta_{q,2}$ by construction of β_q. One can similarly find discrimination by IG and PL.

Now to obtain the degree of voter discrimination, b_{wb}, consider districts of type M where $Q_{R,d} = w$, $Q_{D,d} = b$, and $h_d = logistic(b_{wb} + b + \Delta\xi_{i,d})$. The analogue of equation B.6 for districts of type M is given by:

$$\Rightarrow \log\left(\left(\frac{S_{R,l,d}}{S_{D,l,d}}\right)^{\gamma-1} \cdot \frac{1 - h_d + \beta\sum_l S^{\gamma}_{D,l,d}}{h_d + \beta\sum_k S^{\gamma}_{R,k,d}}\right) = -(\beta_{q,w,l} - \beta_{q,b,l}) - \Delta_i\xi_{i,d} - \Delta_i\xi_{cost,i,l,d} \tag{B.14}$$

Moreover, the equation for uncovering the candidate unobserved valence for districts of type M is given by:

$$\Delta_i\xi_{i,d} = -\mathbb{E}_l\left[\log\left(\left(\frac{S_{R,l,d}}{S_{D,l,d}}\right)^{\gamma-1} \cdot \frac{(1 - P_{R,d}) \cdot \left(1 + \beta\sum_l S^{\gamma}_{D,l,d} + \beta\sum_l S^{\gamma}_{R,l,d}\right)}{P_{R,d} \cdot \left(1 + \beta\sum_l S^{\gamma}_{D,l,d} + \beta\sum_l S^{\gamma}_{R,l,d}\right) - \beta\sum_k S^{\gamma}_{R,k,d} + \beta\sum_k S^{\gamma}_{R,k,d}}\right) + (\beta_{q,w,l} - \beta_{q,b,l})\right] \tag{B.15}$$

Finally, given that we know $\Delta_i\xi_{i,d}$ and recall that we also know β and b_{Rep} (Republican bias) we can recover b_{wb} by re-writing and taking expectations of the M analogue of equation B.8, given by

$$b_{wb} = \mathbb{E}\left[logit\left(P_{R,d} \cdot \left(1 + \beta\sum_l S^{\gamma}_{D,l,d} + \beta\sum_l S^{\gamma}_{R,l,d}\right) - \beta\sum_k S^{\gamma}_{R,k,d}\right) - \Delta_i\xi_{i,d} - b_{Rep} \middle| Q_{R,d} = w, Q_{D,d} = b\right] \tag{B.16}$$

Note, $\mathbb{E}\left[\Delta_i\xi_{i,d} | Q_{R,d} = w, Q_{D,d} = b\right] \neq 0$ and it is not needed as $\Delta\xi_{i,d}$ is recovered using equation B.15 for all districts of type M.

These arguments show that the differences in equilibrium outcomes across districts where both candidates on the ballot are white to those where both candidates on the ballot are black uncovers the degree of election aversion, $\beta_{q,w,C} - \beta_{q,b,C}$. The same argument can be extended for other racial identities and can incorporate gender. The differences in equilibrium outcomes between districts where identity of candidates are distinct to those where it is the same identifies the degree of voter discrimination.[33]

C Quality of GPT-4's prediction of Candidate Race

To assess the accuracy of GPT-4's race predictions, I compare GPT-4's predicted race of House Representatives with the data from CQPress. Race data on GE winners is incomplete for some periods, necessitating the use of GPT-4 to determine the race of congressional candidates. Table A1 reports the proportion of correct predictions by race. At first glance, the "Other" category appears to have lower accuracy. However, further investigation revealed that some candidates were correctly identified by GPT-4 but not by CQPress. Examples include Darren Soto (Puerto Rican father and Italian-American mother, thus of mixed race) and Anna G. Eshoo (Armenian heritage, and therefore should be classified as Middle Eastern). Additional examples exist. Figure 11 presents the confusion matrix comparing GPT-4 and CQPress predictions.

Table A1: Confusion Matrix Table

Ethnicity	Prop. Matched	StdErr
Black	0.904	0.034
Hispanic	0.898	0.039
Other	0.786	0.078
White	0.943	0.008

This table reports the proportion of predictions by GPT-4 on race of candidates that match with race of candidates available on CQPress.

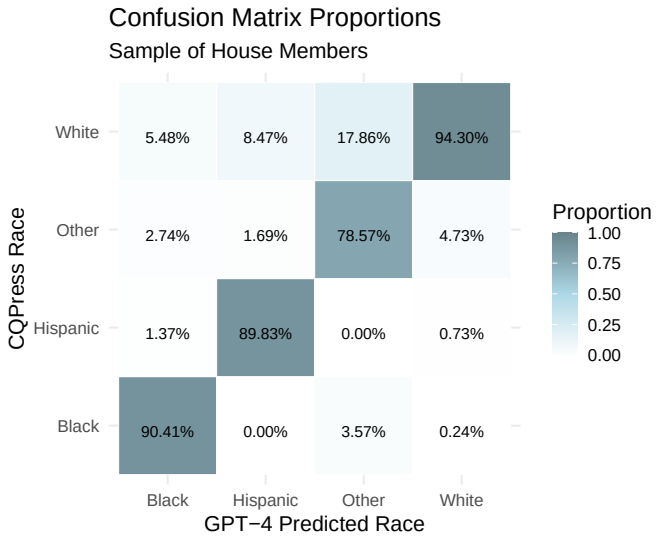
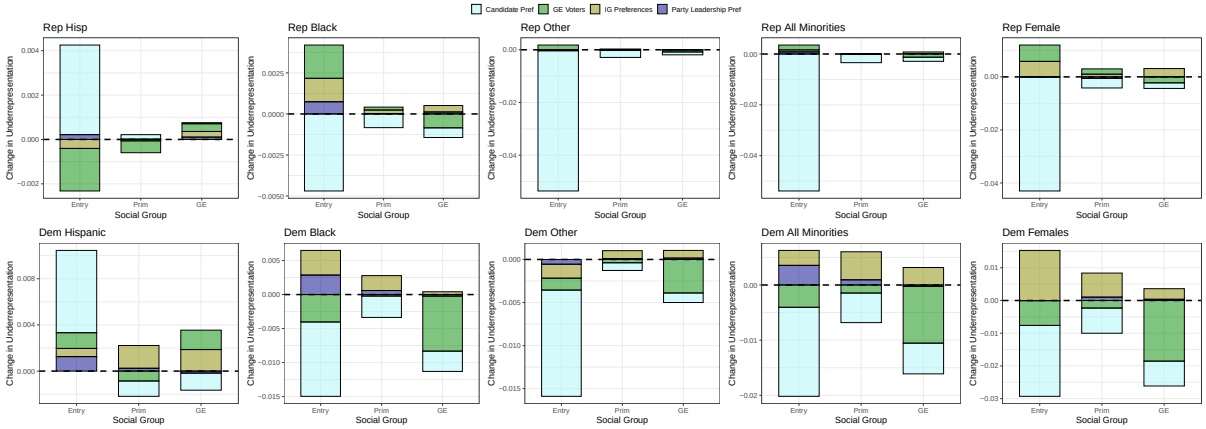


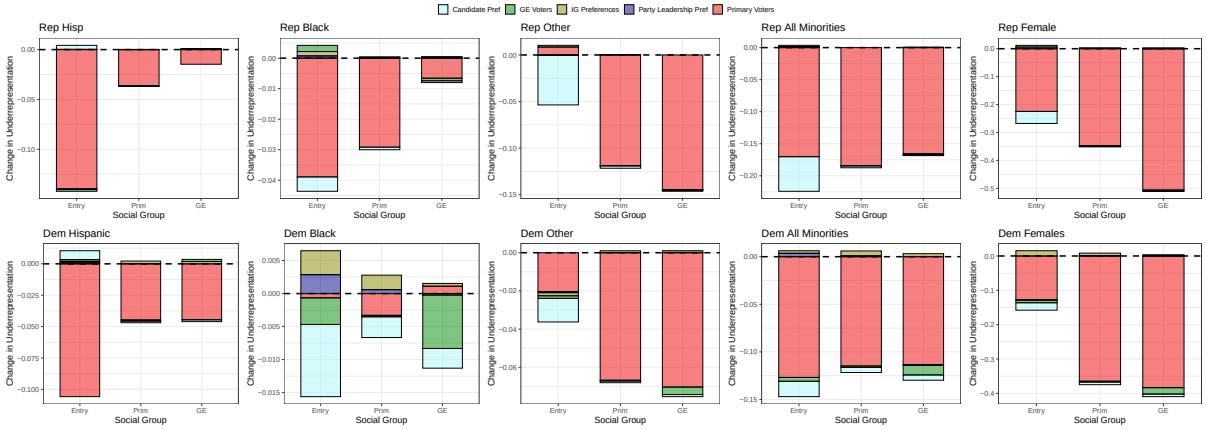
Figure 11: Confusion Matrix Heatmap

[33] As b is expressed in terms of homogenous candidate identity districts.

D Additional Figures



(a)



(b)

Figure 12: This figure shows the changes in underrepresentation when players and contest functions are made indifferent to the race and gender of candidates. Moreover, I refer to "GE contest functions" as "GE Voters" and "Primary contest functions" as "Primary Voters". The y-axis plots change in underrepresentation, defined as $\Delta\text{Underr}_k^{j,l} = (Prop_k - Q_k^{j,l}) - (Propk - Q_k^{j,o})$, where $j \in \{E, PW, GEW\}$ and $k \in$ {Hispanics, Blacks, Other, All Minority races, Female}. $Prop_k$ is the share of social group k in the U.S. population, and $Q_k^{j,l}$ is the predicted share of social group k at stage j (entrants, primary winners, or general election winners) when l is indifferent across race and gender. Here, l indexes the cases when either candidates, interest groups, party leadership, general election voters, or primary voters are indifferent. Panel (a) shows changes for entrants ($j = E$), primary winners ($j = PW$), and general election winners ($j = GEW$) for all l except the case of primary voters. Panel (b) includes primary winners as well. Each subplot is by race to show how the effect of removing taste based discrimination carries on from entry to general election winners.

E Additional Tables

Table A2: General Election Stage Estimates: Controls

Parameter	Estimate	Std. Error	Parameter	Estimate	Std. Error
GE Voter Ideal Point Est			*Conditional Mean Valence*		
$\beta_{I,year}$	0.366***	0.0287	$\beta_{\xi,year}$	-4.56***	0.192
$\beta_{I,\text{Prop. of males}}$	0.318***	0.0204	$\beta_{\xi,\text{Prop. of males}}$	4.31***	0.163
$\beta_{I,\text{median age}}$	-0.171***	0.0212	$\beta_{\xi,\text{median age}}$	-0.0548	0.111
$\beta_{I,\text{median HH income}}$	0.0718	0.0451	$\beta_{\xi,\text{median HH income}}$	-3.33***	0.28
$\beta_{I,\text{Prop unemployed}}$	0.0427	0.0266	$\beta_{\xi,\text{Prop unemployed}}$	-0.553***	0.148
$\beta_{I,\text{less than college}}$	0.395***	0.0407	$\beta_{\xi,\text{less than college}}$	-3.03***	0.245
$\beta_{I,\text{Prop of white}}$	0.73***	0.0375	$\beta_{\xi,\text{Prop of white}}$	-0.583***	0.168
$\beta_{I,\text{Prop of black}}$	0.492***	0.0338	$\beta_{\xi,\text{Prop of black}}$	0.502***	0.189
Observations	4,144				
Penalized Likelihood	34846.64				

*Level of Significance Codes: ***: 0.01, **: 0.05, *: 0.1.* Penalization parameter, λ, is set to 6000. The Monte-Carlo performance of the estimator at this penalization parameter value is reported in Table A3. The table reports the coefficients of controls that govern GE Median Voter Ideal Point, $I_d^G = X_d'\beta_I$, and the mean valence of candidates conditonal on winning a primary congressional district d given by $\mu_{\xi,d} = X_d'\beta_\xi$. The following congressional district characteristics are used: year, proportion of males, median age, median household income, proportion of unemployed, proportion of adults (aged >25) without a college degree, proportion of whites, and proportion of African Americans. The estimates for the main coefficients are reported in Table 4.

Table A3: Monte-Carlo Results for GE Stage

		Mean Estimate			Bias			Mean Squared Error		
Parameter	True Value	D=500	D=1500	D=3000	D=500	D=1500	D=3000	D=500	D=1500	D=3000
β	0.0477	1.37	0.0404	0.0392	-1.32	0.00727	0.00849	58.8	0.00212	0.000356
I_{IG}	0.343	-0.193	0.237	0.551	0.536	0.106	-0.207	1.97	0.755	0.235
$w_{I,PL}$	1.23	1.59	2.28	1.44	-0.364	-1.05	-0.211	3.13	5.35	1.94
$\beta_{I,1}$	-0.6	-0.799	-0.584	-0.629	0.199	-0.016	0.0288	1.75	0.0318	0.013
$\beta_{I,2}$	1.5	1.96	1.56	1.55	-0.462	-0.0587	-0.0478	1.64	0.208	0.0649
$\beta_{I,3}$	-0.6	-0.532	-0.641	-0.633	-0.0678	0.041	0.033	1.29	0.182	0.00891
$\sigma_{cost,C}$	2.72	18.6	2.74	2.8	-15.8	-0.0235	-0.0827	10600	0.245	0.16
$\sigma_{cost,IG}$	2.72	1.75	2.85	2.69	0.97	-0.135	0.0292	4.42	1.37	0.314
$\sigma_{cost,PL}$	2.72	1.78	2.78	2.74	0.937	-0.0575	-0.0264	3.05	0.525	0.303
σ_{ξ}	1.22	14	1.28	1.13	-12.8	-0.0575	0.0928	6230	0.483	0.11
$\beta_{q,C,P}$	0	-0.467	-0.547	0.106	0.467	0.547	-0.106	3.91	2.89	0.437
$\beta_{q,C,white}$	-0.5	-1.39	-1.05	-0.633	0.894	0.553	0.133	9.76	4.1	0.772
$\beta_{q,C,notwhite}$	-1	-1.55	-1.46	-1.21	0.551	0.457	0.209	5.8	2.43	0.586
$\beta_{q,C,male}$	1	0.109	1.29	1.02	0.891	-0.286	-0.0167	13.2	2.53	0.728
$\beta_{q,IG,P}$	0	-1.14	-0.504	0.0295	1.14	0.504	-0.0295	7.58	3.58	0.676
$\beta_{q,IG,white}$	-0.5	-1.55	-1.89	-0.685	1.05	1.39	0.185	8.91	9.93	2.47
$\beta_{q,IG,notwhite}$	-1	-2.07	-1.48	-1.76	1.07	0.48	0.764	5.24	3.14	2.21
$\beta_{q,IG,male}$	1	0.846	0.734	1.13	0.154	0.266	-0.125	7.71	4.26	2.2
$\beta_{q,PL,P}$	0	-0.581	-1.01	-0.079	0.581	1.01	0.079	2.46	3.93	0.876
$\beta_{q,PL,white}$	-0.5	-1.52	-1.02	-0.514	1.02	0.522	0.0142	7.33	2.65	1.22
$\beta_{q,PL,notwhite}$	-1	-1.93	-1.15	-1.38	0.926	0.153	0.379	6.01	3.83	1.32
$\beta_{q,PL,male}$	1	0.613	0.903	0.877	0.387	0.097	0.123	4.5	2.06	0.812
$\beta_{q,V,white}$	-1.5	-2.29	-1.62	-1.73	0.79	0.122	0.233	5.22	0.693	0.279
$\beta_{q,V,male}$	1	1.56	1.39	0.973	-0.557	-0.394	0.0269	2.82	1.25	0.152
$\beta_{q,1}$	-0.6	0.0754	-1.07	-0.587	-0.675	0.469	-0.0129	2.72	1.15	0.129
$\beta_{q,2}$	1.5	1.5	1.99	1.75	0.00397	-0.49	-0.246	2.74	1.17	0.299
$\beta_{q,3}$	-0.6	-0.219	-0.749	-0.749	-0.381	0.149	0.149	4.5	0.413	0.213

Penalization parameter, λ, is set to 6000. Median Voter Ideal Point, $I_d^G = X_d'\beta_I$, and the mean valence of candidates conditonal on winning a primary congressional district d given by $\mu_{\xi,d} = X_d'\beta_\xi$. Two racial groups (white v. non-white) and gender groups (male v. female). Observable candidate and congressional characteristics are not generated but rather randomly picked from the data with replacement for each sample size. These characteristics remain constant across simulation draws. Due to computation constraints, 100-Monte Carlo experiments were performed. Note that mean-squared error for all parameters decreases with the increase in sample-size.

Table A4: Monte-Carlo Results for EP Stage

		Mean Estimate		Bias		Mean Squared Error	
Parameter	True Value	D=1000	D=2000	D=1000	D=2000	D=1000	D=2000
μ_{ξ}	-3	-0.269	-2.73	-2.73	-0.273	7.54	0.142
$\mu_{I,R}$	1	0.304	1.01	0.696	-0.0135	0.485	0.000396
$\mu_{I,D}$	-1	-0.289	-0.991	-0.711	-0.00895	0.51	0.000185
σ_{ξ}	3	1.57	2.85	1.43	0.151	2.08	0.0427
$\sigma_{I,R}$	0.5	0.988	0.504	-0.488	-0.00437	0.238	0.000177
$\sigma_{I,D}$	0.5	0.986	0.476	-0.486	0.0242	0.237	0.00176
δ_{Ideo}	1	-0.0484	0.954	1.05	0.0456	1.11	0.0052
$\delta_{R,hisp}$	-1.5	-1.03	-1.62	-0.472	0.121	0.258	0.0303
$\delta_{R,white}$	0.5	0.202	0.448	0.298	0.0516	0.0991	0.00628
$\delta_{R,black}$	-1.5	-0.841	-1.55	-0.659	0.0466	0.452	0.0117
$\delta_{R,male}$	1.5	0.162	1.52	1.34	-0.0208	1.81	0.00745
$\delta_{D,hisp}$	-1.5	0.108	-1.37	-1.61	-0.133	2.61	0.0729
$\delta_{D,white}$	0.5	0.745	0.539	-0.245	-0.0385	0.0715	0.00829
$\delta_{D,black}$	-1.5	0.19	-1.26	-1.69	-0.243	2.91	0.105
$\delta_{D,male}$	1.5	0.603	1.52	0.897	-0.0196	0.817	0.0292

Median Voter Ideal Points are given by $I_d^G = X'_{d,Prim}\beta_I$. The mean valence of candidates conditional on winning a primary congressional district d given by $\mu_{\xi,d} = X'_d\hat{\beta}_{\xi}$, where $\hat{\beta}_{\xi}$ is the estimate from the GE stage. All racial groups considered for these Monte-Carlo experiments. Observable congressional characteristic (median household income for these MCs) are not generated but rather randomly picked from the data with replacement for each sample size. These characteristics remain constant across simulation draws. Due to computation constraints, 100-Monte Carlo experiments were performed. Note that mean-squared error for all parameters decreases with the increase in sample-size.