Learning and Coordination in the Presidential Primary System*

George Deltas† Helios Herrera‡ Mattias K. Polborn§

November 17, 2015

Abstract

In elections with three or more candidates, coordination among like-minded voters is an important problem. We analyze the trade-off between coordination and learning about candidate quality under different temporal election systems in the context of the U.S. presidential primary system. In our model, candidates with different policy positions and qualities compete for the nomination, and voters are uncertain about the candidates' valence. This setup generates two effects: Vote-splitting (i.e., several candidates in the same policy position compete for the same voter pool) and voter learning (as the results in earlier elections help voters to update their beliefs on candidate quality). Sequential voting minimizes vote-splitting in late districts, but voters may coordinate on a low quality candidate. Using the parameter estimates obtained from all the Democratic and Republican presidential primaries during 2000-2012, we conduct policy experiments such as replacing the current system with a simultaneous system, adopting the reform proposal of the National Association of Secretaries of State, or imposing party rules that lead to candidate withdrawal when pre-specified conditions are met.

JEL Classification Numbers: D72, D60.
Keywords: Voting, Presidential primary elections, Simultaneous versus sequential elections.
JEL Classification Codes: D72, D60.

*We are grateful for the comments of five referees that helped us enormously to improve the paper. We also benefited from comments by seminar participants at Aristotle University, ITAM, Royal Holloway University, Universidad Carlos III, University of Lancaster, University of Leicester, University of Missouri, Texas Tech, Toulouse School of Economics, the Midwest Political Science Meetings, the Econometric Society Summer meetings (St Louis), CRETE 2011 (Greece), as well as Costas Arkolakis, Simon Anderson, Sofronis Clerides, Andrei Gomberg, Emilio Gutierrez, Brian Knight, Roger Koenker, and Rainer Schwabe for helpful comments. Financial support from the Spanish Ministry of Science (ECO2008-01300; Deltas) and from National Science Foundation Grant SES-1261016 (Polborn) is gratefully acknowledged. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation or any other organization.

†Department of Economics, University of Illinois, 1407 W. Gregory, Urbana, IL 61801 USA, E-mail: deltas@uiuc.edu
‡Department of Applied Economics, HEC Montreal. Email: helios.herrera@hec.ca
§Corresponding author. Department of Economics and Department of Political Science, University of Illinois, 1407 W. Gregory, Urbana, IL 61801 USA, E-mail: polborn@uiuc.edu.

1 Introduction

A fundamental question in the analysis of politics is how institutions influence election results and policy outcomes. In all elections with three or more candidates, coordination is an important issue for like-minded voters because miscoordination can lead to the electoral victory of a candidate disliked by the majority. Moreover, when the quality of candidates is uncertain and learned only imperfectly by voters over the course of the campaign, an additional problem is that strategies that foster immediate coordination among like-minded voters may limit the extent of information aggregation that can be achieved.

As a practical example in which these issues are of first order importance, consider the new "top-two primary system" in California in which a – possibly large – number of candidates from all parties jointly run at the primary stage, and the top two vote getters proceed to the general election runoff stage. At the primary stage, Democrats and Republicans, want to make sure that the "best" candidate from their respective parties proceeds to the general election. However, there is a lot that can go wrong if it is not obvious who that is. Consider, for example, the 2012 primary in California's 31st Congressional district, a liberal-leaning district with a Hispanic majority. Four Democratic candidates split the votes of liberal-leaning voters among each other, while there were only two Republican candidates, both of whom obtained more votes than the top Democrat. Consequently, the two Republicans proceeded to the general election, and as a result of this coordination failure among Democratic voters, a Republican won. Interestingly, Democrats almost repeated this coordination failure in 2014, when the second-strongest Republican came within one percentage point of the strongest Democrat. Conversely, though less conspicuously, the fear of coordination failure may sometimes lead to premature coordination on a suboptimal candidate.

Empirically analyzing the effects of voter learning and coordination in elections, while crucial for the optimal design of political institutions, is usually exceedingly difficult because we only observe voters make a decision once, i.e. when they vote at the end of the campaign. Voter learning and/or coordination may be facilitated by campaign advertising, media, or word-of-mouth, but these dynamics are hard to observe in a way that would lend itself to a quantitative analysis. However, there is one institution in which voting occurs sequentially and which is therefore uniquely suited to study these issues: The U.S. Presidential Primary system provides a natural experiment case study for learning and coordination during an election campaign,[1] but our findings carry broader implications on the extent to which lack of coordination can sway election outcomes in other simultaneous or two-round election contests.

An analysis of the primary system is also substantively important, both because it is a crucial aspect of the selection for the most powerful office in the world, and because the structure of the primary system has actually changed substantially in the past and is likely to continue to be modified in the future. The nomination process is one of the most controversial institutions of America's contemporary political landscape: Its sequential structure is perceived as shifting disproportionate power to voters in early primary states, and many states have shifted their primaries earlier and earlier over the last several election cycles, while the national parties have tried to steer against this movement, by punishing states that go "too early."[2]

The main alternatives to the current status quo of a sequential system are a nationwide primary to be held on the same day, and a proposal by the National Association of the Secretaries of State (NASS) for regional primaries.

[1] Different states have their presidential nomination elections organized as either primaries or caucuses. Since we are only interested in the temporal organization of the entire nomination process, we will, in a slight abuse of terminology, call all of these contests "primaries."

[2] For example, the Economist (Feb. 15, 2007) describes the efforts of the national parties to maintain the sequential nature of the primary process: "The national Democratic and Republican parties are not pleased. They have tried to bribe states by promising them more delegates (and thus a bigger say in the nomination process) if they hold their elections later, and have threatened to deprive them of delegates if they insist on moving.

According to the NASS proposal (see Stimson (2008)), Iowa and New Hampshire would always vote first, followed by four regional primaries (for the East, Midwest, South and West regions) scheduled on the first Tuesday in March, April, May or June of presidential election years. The sequence of the four regions would rotate over a 16-year cycle. In our framework, we can analyze (i) under which circumstances the temporal organization makes a difference for who wins the nomination, and (ii) whether such a change is beneficial for voters from an ex-ante perspective.

To better understand the trade-off between learning about candidate quality and voter co-ordination, consider the following - half-fictional - example of a nomination contest. At the time of the first elections, three serious contenders whom we call C, E and O, competed for the nomination. These candidates differ in some characteristics that are relevant for voters. First, candidate C has experience in Washington and in the White House, while candidates E and O run as "Washington outsiders" or "change candidates".

Suppose that, ceteris paribus, some voters prefer a candidate with Washington experience, while others (the "change voters") prefer an outsider with a strong grass-roots base. In addition, there is uncertainty about the valence of each candidate. If the primary elections were to take place simultaneously in all states, then it is quite plausible that C wins most states, as E and O split the change voters. In contrast, in a sequential system, "change" voters in states that hold their primaries after the first ones can observe the early election results and vote accordingly; also, in expectation of such coordination, the trailing candidate may drop out early. For example, if O gets more votes than E in the early elections, E will drop out realizing that even voters with ranking $E \succ O \succ C$ may in the remainder of the primary vote for O, because they have determined that E has no chance of winning and they prefer O among the remaining candidates. In this case, O will win the nomination if a majority of the electorate prefers him to C.

Such voter migrations between candidates may be crucial for election outcomes. In the 2008 Democratic nomination contest for instance, the majority of Edwards supporters appears to have migrated to Obama after Edwards dropped out of the race, and Obama would likely not have won the nomination had the primary election been held simultaneously on all states.[3] Arguably, analogous effects have arisen in the 2012 Republican nomination contest, with Rick Santorum and Newt Gingrich vying for recognition as the "conservative challenger" of the "establishment candidate" Mitt Romney and splitting much of the conservative vote for as long as both stayed in the race.

The benefit of a sequential system in our example is that, in most districts, the change voters do not split their votes, thus increasing the likelihood that a change candidate wins. There is, however, also a disadvantage when voters are uncertain about candidate valences: Conditioning coordination on only one or few initial elections raises the possibility that the weaker change candidate comes out on top, and if such an early electoral mistake occurs, it cannot be corrected in the remaining districts. Learning about candidate quality is a very relevant problem in presidential primaries: While most candidates are accomplished politicians such as governors or members of Congress, very few of them are already household names for a national audience.[4] The objective of our model is to provide a framework for the analysis of the trade-off between coordination and voter learning.

The paper is organized as follows. Our theoretical model in Section 2, develops the simplest framework in which

[3]For example, Moulitsas (2008a) cites a Rasmussen poll for Missouri from January 31 (the last one conducted with Edwards in the mix) before the primary one week later. The preference numbers in the Rasmussen poll were Clinton 47, Obama 38, Edwards 11, while the actual election results were Obama 49.3, Clinton 47.1, Edwards 1.7. Similarly, in a 12/26-30, 2007 poll by Opinion Research Corp for CNN (cited by Moulitsas (2008b)), 36% of Iowa Democrats polled declare that Edwards was their second choice, 25% name Obama, but only 11% name Clinton as their second choice. Since all three candidates were very close in terms of *first* preferences, this suggests that most Obama and Edwards supporters had the respective other candidate as their second preference.

[4]Moreover, in addition to past achievements, voters also care about how candidates acquit themselves under the pressure of an intense campaign under the spotlight of the national media. Thus, learning about candidate quality naturally proceeds throughout the entire primary process.

the issues of learning and coordination interact with each other, and provides some guidance as to which factors affect this trade-off. The net effect can go in either direction, so that the question of the optimal voting system is a quantitative one. In Section 3, we estimate the structural parameters of our theoretical model using data from the 2000, 2004 and 2008 Democratic primaries and separately for the 2000, 2008 and 2012 Republican primaries. The estimated parameter values show that both key features of the theory – voter learning about candidate valence, and unequal substitutability of candidates with different political positions – are quantitatively important. The main purpose of the estimation is not to "test" the model in a classical sense but rather to develop reasonable starting values for our institutional simulations.

The simulations in Section 4 are the core substantive results. All simulations consider races with three candidates competing for the nomination, two of whom share the same political position. We compute the distribution of election outcomes under several different sequencing scenarios, and different assumptions about the dropout decision of the weakest candidate under sequential voting. In terms of sequencing scenarios, we look at simultaneous voting in all states; completely sequential voting; and the NASS proposal of regional primaries (i.e., two sequential single-state votes, followed by 4 rounds, in each of which 12 states vote simultaneously).

Our results show that a sequential election results in the highest expected valence and the highest probability that the Condorcet winner is elected. The NASS proposal comes in as a very close second to sequential primaries (slightly worse in the baseline case, but slightly better in some sensitivity analysis variations). In contrast, a simultaneous election performs substantially worse. These results do not just hold for the baseline cases using the point estimates of parameter values estimated from the Democratic and Republican primaries, but also appear to be very robust with respect to reasonable variation in the parameter values. The only case in which the performance of simultaneous primaries comes close to sequential ones is when the importance of position differences between candidates – and thus of vote splitting, when there is an unequal number of candidates in both positions – is much smaller than the estimated parameter for both parties.

We also consider variations in dropout timing, relative to the dropout threshold function estimated from the 2000-2008 Democratic primaries. We find that, from a social point of view, both Democratic and Republican candidates drop out too late, especially the former. This is an indication that individual candidates do not internalize the negative externality that their continuing campaign imposes on their party and its voters. It presents a opportunity for intervention by the Parties to nudge candidates into withdrawing earlier than they currently do.

Our paper contributes to the literature on the classical coordination problem of a divided majority in plurality rule elections. As Bouton and Castanheira (2012) point out, this is both a very relevant problem in many elections (e.g., the 2000 U.S. Presidential election, with Gore and Nader splitting the liberal vote), and one that has spawned a large number of suggested reforms of the voting system to deal with the problem (such as approval voting in Bouton and Castanheira (2012)).

In the political science literature, the study closest to our focus on the role of early primaries as a coordination device is Bartels (1987), who analyzes the 1984 Democratic presidential primary and describes the coordination process of the Democratic voters (i.e., Hart versus other non-Mondale candidates).[5]

Knight and Schiff (2010), provide both a theoretical model and an empirical study of the 2004 Democratic primary. In contrast to our model, though, their model is not designed to analyze the optimality of different temporal structures of the primary process, and also does not have a trade-off between coordination and learning. Knight and Hummel

[5]Other studies analyzing similar relationships include Bartels (1985) for the 1980 Democratic primaries and Kenny and Rice (1994) for the 1988 Republican primary, but all of these focus implicitly on a two-candidate framework.

(2015) extend that model to analyze the welfare effects of sequential and simultaneous primaries. They do not allow for horizontal differentiation between candidates (i.e., voters do not have policy preferences over candidates), but allow for candidate quality being drawn from different ex-ante distributions. We discuss the differences between their paper and ours in more detail in Section 4.

The vast majority of election system reform research is either entirely theoretical (i.e., postulates a certain setting and then analyzes how different election systems perform in that setting) or experimental. Most of the theoretical literature on sequential primaries has focused on elections with two alternatives (see, e.g., Dekel and Piccione 2000, Callander 2007, Ali and Kartik 2012). With just two candidates, the problem of coordination does not arise. Moreover, these models are primarily positive in nature and do not focus on optimal institutional design (exceptions to this, but still in a two-candidate framework are Klumpp and Polborn (2006) and Schwabe (2010)).[6]

In a clever lab experiment, Morton and Williams (1999, 2001) analyze the trade-off between learning and coordination in simultaneous and sequential elections, and show that both effects occur in later elections in their experiment. Our paper builds upon theirs in that we take it as given that voters in later elections learn about candidate quality and try to coordinate with other voters. Bouton, Castanheira, and Llorente-Saguer (2012) analyze a problem of learning and coordination in a setting with three candidates and repeated elections. While their theory does not have much power in terms of predicting whether coordination will occur, they also run an experiment in which those voters who face a coordination problem indeed use the outcome of the first election as a coordination device and ignore information that they receive in later rounds. Their result thus shows the significance of the trade-off between learning and coordination that is at the core of our paper.

Our main value added relative to these papers is that we analyze this trade-off using data from actual elections. Because the optimal primary structure depends on the size and the interaction of the two effects in a nontrivial way, policy implications for optimal institutional design should be based on data derived from real-world primaries, rather than on laboratory experiments.

2 Model

We now introduce the model and provide theoretical results that demonstrate the tradeoff between learning and coordination for a special case. In the following sections, we analyze the general model empirically.

2.1 Setup

The set of states (i.e., electoral districts) is $\{1, \ldots, S\}$, with typical state s. There is a set of J candidates who differ along two dimensions. First, the parameter v_j measures Candidate j's valence which is a common value characteristic appreciated by all voters, like competence. Second, candidates are also horizontally differentiated: they have either position 0 or 1 on a binary policy issue, and, without loss of generality, the first j_0 candidates are fixed at $a_j = 0$, while the other $j_1 = J - j_0$ candidates are fixed at $a_j = 1$.

Voter i's utility from a victory of Candidate j is

$$U_j^i = v_j - \lambda |a_j - \theta^i| + \varepsilon_j^i, \tag{1}$$

[6] Another noteworthy contribution in this area is Anderson and Meagher (2011) who embed the primary election into a framework that includes party formation and party competition.

where θ^i is voter i's preferred position on the policy issue, and λ parameterizes the importance of the policy issue relative to valence. The last term in the utility, ε^i_j, drawn from $N(0, {\sigma_\varepsilon}^2)$, is an individual preference shock of voter i for Candidate j, as in probabilistic voting models.[7]

While the ε preference shocks are distributed identically for all candidates, voters may have asymmetric preferences for policy positions. Specifically, the proportion of the electorate in state s with preference for $a = 1$ is denoted μ^s and is common knowledge for all players.

The policy dimension is known to voters and captures the notion that some candidates are similar to each other and hence close substitutes for most voters, while there is a more substantial difference to some other candidates (see e.g. Krasa and Polborn (2010)). The dimension does not need to be interpreted as "policy" in a narrow sense; for example, in the 2008 Democratic primary, voters may have different views on the desirability of "political dynasties" (see Dal Bo, Dal Bo, and Snyder (2009)).

In contrast, the valence dimension is unknown to voters. From their perspective, each candidate's valence is an independent draw from a normal distribution $N(0, {\sigma_v}^2)$. Voters cannot observe v_j directly: in state s they observe a signal $Z^s_j = v_j + \eta^s_j$, where the additional term for Candidate j, η^s_j, is an independent draw from a normal distribution $N(0, {\sigma_\eta}^2)$. η^s_j is a state-specific (as opposed to voter-specific) observation error: voters in the same state receive their news about the candidates from the same local news sources so that errors, if any, are not individual-specific.[8]

Given their own signal, and possibly the election results in earlier states (from which the signals in those earlier states can be inferred), voters rationally update their belief about the valence of candidates. Let $\hat{v}^s_j$ denote the valence of Candidate j that is expected by voters in district s.

If voting is sequential, it may be the case that the set of candidates shrinks over time, and we let J^t be the set of "relevant" candidates in period t elections and assume that each voter votes "sincerely," i.e. for his most preferred candidate from this set.[9] That is, voter i in district s (which votes at time t) votes for Candidate j if and only if

$$j \in \arg\max_{j' \in J^t} (\hat{v}^s_{j'} - \lambda |a_{j'} - \theta^i| + \varepsilon^i_{j'}).^{10} \tag{2}$$

2.2 Coordination and learning in a special case

We now turn to a special case in which we can illustrate the effects of voter coordination and voter learning about candidate quality. There are $J = 3$ candidates, with positions $a_1 = 0$ and $a_2 = a_3 = 1$. Assume that $\lambda \to \infty$, i.e., voters value candidate positions much more than valence v and idiosyncratic preferences ε. Thus, all voters with preferred position $\theta^i = 0$ vote for Candidate 1, while those voters with $\theta^i = 1$ either vote for Candidate 2 or Candidate 3.

[7] A possible interpretation of this term is that candidates also differ in a large number of other dimensions for which voters have different preferences. The policy dimension modeled explicitly ($a_j = 0$ or $a_j = 1$) should then be understood as the most important dimension in this example. See, e.g., Lindbeck and Weibull (1987), Coughlin (1992) or Persson and Tabellini (2000) for a review of the probabilistic voting literature.

[8] Of course, in reality, there are plausibly both common and idiosyncratic observation errors. To simplify the model and gain some tractability, we focus on the state-specific observation error. If, instead, observation error terms were only individual-specific, then the true valence of candidates would be perfectly known after the election results of the first state, which appears unrealistic.

[9] In elections with more than two candidates, there are generally many Nash equilibria in undominated strategies. However, sincere voting is a standard assumption in the literature for multicandidate elections, and also appears to capture voter behavior in many elections (see Degan and Merlo (2006)). Also note that in our model participation is exogenous and thus we do model how changes in the field of candidates affects incentives to vote and voting turnout. For some recent work on voter turnout in the US Presidential primaries see Kawai and Watanabe (2010).

[10] Since the distribution of ε is continuous, the measure of voters who are indifferent between 2 or more candidates is equal to zero, so it is irrelevant for the election outcome how those voters behave.

Furthermore, we assume that the proportion of the total population with preference for $a = 1$ is equal to μ in all electoral districts ($\mu^1 = \mu^2 = \ldots = \mu^N \equiv \mu$).

Clearly, if $\mu < 1/2$, then Candidate 1 is the Condorcet winner, and his supporters form a majority in each district. If $\mu > 1/2$, then either Candidate 2 or Candidate 3 is the (full information) Condorcet winner, depending on which one of them has the higher valence. This generates the basic coordination problem for voters whose preferred position is 1: if candidates 2 and 3 split the votes of voters who prefer position 1, then Candidate 1 may win even if he is not the Condorcet winner (i.e., the candidate who would be preferred by a majority of voters to all other candidates, if valences were known).

We analyze two particular temporal organizations of the primary system. Under simultaneous elections, all S states vote at the same time. Under sequential elections, one state votes at $t = 0$, and the remaining $S - 1$ states vote at $t = 1$, after observing the election outcome in the first state. We assume the states to have the same size and the number of states, S, to be large enough so that the vote outcome at $t = 1$ determines the overall result of the sequential election. However, the outcome of the first election at $t = 0$ is used by later voters as a coordination device: The set of relevant candidates at $t = 1$ is formed by excluding either Candidate 2 or 3 (i.e., one of the two candidates in position 1), depending on who did worse in the first state.

Proposition 1, proved formally in the Appendix, characterizes the equilibrium for the two different primary systems. By Condorcet loser, we mean the candidate who would lose against either opponent.

Proposition 1 *Assume that Candidate 1's policy position is 0 and both Candidate 2 and 3 have policy position 1, and that λ is large relative to σ_v and σ_ε.*

If $\mu < 1/2$, Candidate 1 is the Condorcet winner. If $1/2 < \mu < 2/3$, Candidate 1 is the Condorcet loser, and the candidate with the higher valence among Candidates 2 and 3 is the Condorcet winner.

1. *If $\mu < 1/2$, Candidate 1 wins under both a simultaneous and a sequential primary system.*
2. *If $1/2 < \mu < 2/3$,*
   (a) *In a sequential primary system, either Candidate 2 or Candidate 3 wins. The probability that the Condorcet winner wins is decreasing in σ_η and increasing in σ_v.*
   (b) *In a simultaneous primary system, either the Condorcet winner or Candidate 1 wins. There exists $\mu^* \in (1/2, 2/3)$ such that Candidate 1 (the Condorcet loser) wins the nomination with positive probability for every $\mu < \mu^*$.*
3. *If $\mu > 2/3$,*
   (a) *In a sequential primary system, Candidates 2 and 3 each win with positive probability, while Candidate 1 cannot win.*
   (b) *In a simultaneous primary system, the Condorcet winner wins with probability 1.*

We now discuss the intuition for these results. First, if $1 - \mu > 1/2$, Candidate 1 receives an absolute majority of votes in every district, whether he competes against one or two opponents. The election system only affects whether the votes of type $\theta = 1$ voters are split or united, but even perfect coordination cannot change that Candidate 1 wins.

If $\mu \in (1/2, 2/3)$, type 1 voters are in the majority, and thus either Candidate 2 or Candidate 3 is the Condorcet winner. However, since Candidate 1 receives more than one-third of the votes, it is possible that he receives a plurality

in some or all districts. In this case, interesting differences between sequential and simultaneous primary systems arise. The advantage of a sequential system is that it avoids vote splitting and thus prevents a victory of the Condorcet loser; however, the winning candidate may be of lower quality than the candidate who dropped out. In contrast, in a simultaneous election system, the law of large numbers guarantees that the better of Candidates 2 and 3 wins more votes than the weaker one. However, since there is vote splitting between Candidates 2 and 3, Candidate 1 may win even though he is the Condorcet loser.

To see these effects in more detail, consider first sequential elections. Since $\mu > 1/2$, either Candidate 2 and Candidate 3 (whoever wins more votes in the first district) will win all remaining districts. Thus, in a sequential organization of primaries, it is impossible that the Condorcet loser wins. However, because the signal of first-district voters is not perfect, the Condorcet winner may fare worse in the first district than his competitor with the same position, and in that case, the error cannot be corrected in the second period.

An explicit formula for the Condorcet winner's winning probability is derived in the proof. Intuitively, a higher σ_η means that there is a larger chance that the difference of observation mistakes for the two candidates outweighs their valence difference, so that voters in the first district mistakenly perceive the worse candidate as the better one. In contrast, if σ_v increases, this increases the expected valence difference between the better and the worse candidate and thus raises the probability that the Condorcet winner wins.

Now consider simultaneous elections when $\mu \in (1/2, 2/3)$. Since Candidate 1's vote share, $1-\mu$, is larger than $\mu/2$, it is possible that voters with a preference for Candidate 2 or 3 split in such a way in a district that Candidate 1 wins a plurality. How often this happens depends on parameters. If there is a large difference between the perceived valences of Candidates 2 and 3, and if the idiosyncratic preference differences captured by ε are sufficiently small for most voters, then almost all of them agree on one candidate, and vote splitting is minimal. In these cases, the Condorcet winner is likely to win a plurality. In contrast, if perceived valence differences between candidates are small or idiosyncratic preference shocks are large, then both Candidate 2 and 3 receive a substantial fraction of support, and Candidate 1 may win.

If $\mu > 2/3$, type 1 voters are in the majority, and thus either Candidate 2 or 3 is the Condorcet winner. In contrast to the case that $\mu \in (1/2, 2/3)$, though, Candidate 1 cannot win since the electorate's preference distribution is sufficiently extreme to make up for any extent of vote splitting between Candidates 2 and 3. In a simultaneous elections system, the law of large numbers guarantees that the better candidate (among Candidates 2 and 3) wins a larger number of districts than his weaker competitor. Thus, when $\mu > 2/3$, the Condorcet winner always wins under simultaneous elections. In contrast, in a sequential election system, there can still be mis-coordination on the worse candidate among Candidates 2 and 3 because, depending on the outcome of the first district, the Condorcet winner may be eliminated.

2.3 Learning from vote shares in sequential elections

We now turn to the theoretical foundations of voter updating about candidate valence and vote-share determination for the empirical analysis. An overview is presented here, the formal details are in the Appendix. We first analyze how the candidates' vote shares are determined by fundamentals and voter beliefs about candidates, then turn to the development of beliefs given the signals observed over the course of the campaign.

Suppose that the beliefs of voters in district s are given by the vector $\hat{v}^s = (\hat{v}^s_1, \hat{v}^s_2, \ldots, \hat{v}^s_J)$. Beliefs about candidate valence, together with an individual's idiosyncratic preferences, determine the candidate he votes for. In particular, a

voter of type θ votes for Candidate $j \in J_0^s$ if and only if, for all $j' \neq j$,

$$\hat{v}_j^s + \varepsilon_j - \lambda d(j,\theta) > \hat{v}_{j'}^s + \varepsilon_{j'} - \lambda d(j',\theta), \tag{3}$$

where $d(j,\theta)$ measures the distance between Candidate j and voter type θ (i.e., $d = 0$ if voter type and candidate agree, and $d = 1$ when they disagree).

In the Appendix, we show how integrating over all possible realizations of idiosyncratic preferences provides us with a system of vote share equations, (26), which are functions of the valence beliefs $\{\hat{v}_1^s, \hat{v}_2^s, \ldots\}$. Proposition 2 Corollary 1 in the Appendix show that this equation system has a solution such that observing the outcome in state s allows voters in later states to essentially recover the estimated vector of candidate valences in state s, and thus the valence signals Z_j^s.

Note that vote shares are determined only by the *difference* between the candidates' estimated valences (cf. equation (3)), so we can only determine those differences. However, it is also immaterial which of these possible beliefs a voter in a later state uses to infer the signals observed by the voters of that state.[11]

We now turn to the process of updating a voter's belief. If a voter has an ex-ante belief (i.e., before seeing his own state-specific signal) about candidate j's valence distributed according to $N(\hat{v}_{j,s_0}, \sigma^2_{j,s_0})$ and receives a state-specific signal Z_j^s, then the candidate's ex-post valence is normally distributed with expected value

$$\hat{v}_j^s = \frac{\sigma_\eta^2}{\sigma^2_{j,s_0} + \sigma_\eta^2}\hat{v}_{j,s_0} + \frac{\sigma^2_{j,s_0}}{\sigma^2_{j,s_0} + \sigma_\eta^2}Z_j^s \tag{4}$$

and variance

$$(\sigma^s_{v_j})^2 = \frac{\sigma^2_{j,s_0}\sigma_\eta^2}{\sigma^2_{j,s_0} + \sigma_\eta^2}. \tag{5}$$

In the first state(s), initial beliefs are evidently given by the fact that candidate valences are drawn from $N(0, \sigma_v^2)$. Applying (5) recursively[12] shows that the coefficient of the candidate valence signal in state j in (4) takes the same value for all candidates. Thus, an increase in the values of all valence signals by a constant increases ex-post valence estimates of all candidates by the same amount. Since vote shares are determined by differences in ex-post valences, they are unaffected. Therefore, signal realizations can be normalized by subtracting a constant so that the signal of the first candidate is equal to zero.

3 Empirical analysis of the 2000–2012 Presidential primaries

We now turn to the empirical analysis. Our ultimate objective is not primarily to test our theoretical model for these particular primary races, but rather to obtain roughly plausible values for parameters on which we can base simulations of the effects of different primary structures. Using the point estimates as a starting point, we then analyze the robustness of the results to changes in parameters.

[11] By observing vote shares in the election of a prior state, a voter can infer beliefs about valence up to a common constant for all candidates. Voters determine their preferred candidate on the basis of differences in ex-post perceived valence, and these differences are determined by differences in the valence signals observed by voters of the state. In other words, a uniform shift of the ex-ante beliefs about all candidates by c translates into a uniform shift of the ex-post beliefs (i.e., after the state-specific signal), leaving the difference between the valence estimates for the different candidates, and hence the voter's voting decision, unaffected.

[12] Note that the application of (4) and (5) is by round, i.e., all states voting in a particular round use values of $\hat{v}_{j,s_0}$ and σ^2_{j,s_0} as obtained from the signals up to the end of the previous round.

3.1 Data

Our dataset consists of the vote shares from the 2000-2012 contested U.S. Presidential primaries (i.e., we exclude the 2004 Republican and the 2012 Democratic primaries in which there was no serious challenge to a sitting U.S. President). In the Democratic primaries, we include Al Gore and Bill Bradley for 2000, John Kerry, John Edwards, Howard Dean, Wesley Clark and Joe Lieberman for 2004, and Barack Obama, Hillary Clinton, and John Edwards for 2008. In the Republican primaries, we include John McCain and George Bush for 2000, John McCain, Mitt Romney, and Mike Huckabee for 2008, and Mitt Romney, Newt Gingrich and Rick Santorum for 2012. In all years, we thus exclude some minor candidates who were generally expected to have a negligible chance of winning the nomination.[13] We generally include primary vote shares from all states and the District of Columbia, while we exclude primaries in U.S. overseas territories (e.g., Guam). We also exclude the Michigan 2008 primary because it was held earlier than allowed for by Democratic party rules, and the names of Obama and Edwards were not on the ballot there.

A key component of the model is that candidates are distinguished by their horizontal position. For the Democrats, we have presented evidence in the introduction that voters in the 2008 contest viewed Edwards and Obama as relatively close substitutes for each other, while Clinton is farther away. There are certainly different potential explanations for why this was the case, and which one applies is immaterial for our estimation. Our preferred interpretation is that Obama and Edwards were perceived as outsiders, while Hillary Clinton was seen as part of the Democratic establishment and representing a continuation of the political philosophy of her husband's administration.[14] Voters may have different views on the desirability of such political dynasties (Dal Bo, Dal Bo, and Snyder (2009) document the importance of family connections for political careers in the U.S. Congress).

In 2004, we group John Kerry and Joe Lieberman together as "establishment candidates," and John Edwards, Howard Dean and Wesley Clark as "outsider candidates." In 2000, Gore is the establishment candidate and Bradley the outsider. These assignments appear to be reasonable summaries of the perception of the candidates by political experts, the media and the candidates themselves.[15]

For the Republicans, the main political divide is between "moderates" and "conservatives." In the 2000 primary, McCain was the moderate candidate, while Bush relied on the conservative base (before somewhat re-positioning in the general election). In the 2008 primary, McCain was again the moderate candidate, and both Romney and Huckabee ran as conservatives. It is telling that Romney recognized that he and Huckabee were competing for the same pool of candidates, and suggested that one of them withdraws so as to increase the chances of the other to succeed in beating McCain. When Huckabee refused to withdraw, Romney did so himself. In the 2012 primary, however, Romney was the moderate standard-bearer, facing Gingrich and Santorum who were supported by the Republican base. Their splitting of the conservative vote helped him win the nomination. For the candidates included in our analysis, we obtain the vote percentage in the primary or caucus of each state from the Federal Election Commission, supplementing from

[13] In 2000 we exclude Keyes, whose run was more of a publicity stunt. In 2004, we exclude Kucinich and Sharpton who were always considered protest candidates; Gephardt is dropped because he only competes in Iowa, which was skipped by two major candidates and is therefore not included in the estimation. In 2008, Guiliani and Thomson run unconventional campaigns effectively skipping most of the first three rounds; thus, we start the estimation after their withdrawal (recognizing that informative signals have been received in the early rounds). In 2012 we exclude Paul who was more of a Libertarian than a mainstream Republican candidate.

[14] Deltas and Polborn (2015) argue that the single most salient partition of the Democratic candidates between in the three last presidential primaries was whether a candidate is perceived to be an insider of the Washington establishment, or rather draws his strength from the grass roots, and runs as an "outsider." In contrast, the liberal versus moderate distinction appears to be of lesser importance.

[15] See, e.g., `http://en.wikipedia.org/wiki/Democratic_Party_(United_States)_presidential_primaries,_2000` for a description of the candidates in the 2000 campaign, and analogous for the other years.

other sources as needed. We rescale the data such that the vote shares add up to 100%.[16]

Before turning to identification and formal empirical analysis, it is useful to provide some motivation using key stylized facts from the three Democratic primaries (similar facts can be reported for the Republican primaries, but are omitted to save space). These data features underlie many of the moments used in the estimation of the structural model. The 2000 primary involved only two candidates, and thus does not illuminate substitutability between the candidates of different positions. It does, however, illuminate the reduction in vote share variability as the primary progresses, a key prediction of Bayesian learning of candidate valence. The vote shares of the candidates were broadly constant over the entire primary: The (unweighted) Gore share in the first 5 rounds, consisting of 7 states, was 66.5%, and dropped slightly to 63.8% in round 6 consisting of 8 states. The standard deviation of vote shares, however, dropped substantially from 12.3% to 6.9%.

The 2004 primary, where there were five serious candidates competing, provides a good illustration of differences in substitutability between candidates of the same political position and substitutability of candidates in different political positions. In election rounds 4 to 7, consisting of 7 states, three outsider candidates were competing against Kerry (an "insider" candidate). For the last two rounds, consisting of 13 states, the surviving outsider candidate, Edwards, competed against Kerry. The vote share of Edwards in the first group of states was 19.5%, increasing to 29.6% in the second group of states.[17] The corresponding figures for Kerry are 54.0% and 70.4%. Even though Kerry won more votes than Edwards in absolute number, his initial percentage was much higher reflecting a higher valence. A more appropriate comparison is the percentage increase of the two candidates' vote shares. That of Edwards increased by 52 percent, while that of Kerry by only 30 percent. This provides a clear indication that Edwards was a closer substitute to Dean and Clark than is Kerry. The 2004 primary cannot easily illuminate changes in vote share variability since there are no long stretches of the primary contest where the set of candidates is fixed.

The 2008 primary is a classic illustration of a both learning and coordination in a presidential primary. Figure 1 plots the mean vote shares of the three candidates for the first five states, the group of Super Tuesday states, and post-Super Tuesday states. It also plots the standard deviation of Clinton's vote shares in the latter two groups.

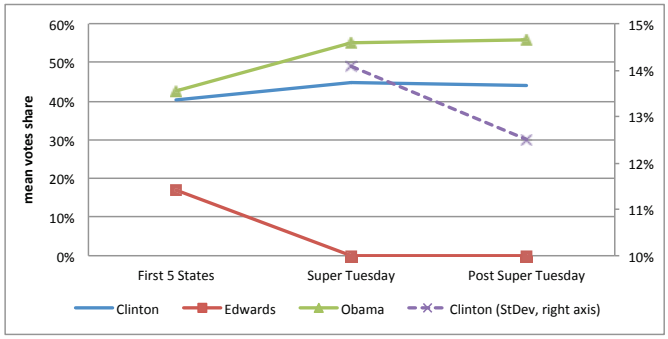

Figure 1: 2008 Democratic primary vote shares

From the mean vote share data, it is apparent that Obama's vote share increased by far more than Clinton's vote share following the withdrawal of Edwards (after the first five states, before Super Tuesday). This suggests that most voters who ranked Edwards first preferred Obama to Clinton, i.e., that Obama and Edwards are closer substitutes than

[16] These data, along with the information about the round in which each state voted, are available in the supplementary files.

[17] Vermont is excluded from all these figures because Dean won it despite having withdrawn earlier from the primary.

Clinton and Edwards. Notice that, after Edwards supporters have reassigned themselves to one of the two remaining candidates, there is no general trend in the vote shares: The vote shares of Obama and Clinton for the Super Tuesday states are almost identical to their vote shares post-Super Tuesday. The interpretation of the decline of the standard deviation of Clinton's vote share post-Super Tuesday relative to the Super Tuesday states is straightforward. The voters' priors about the candidates are less firm on Super Tuesday than following Super Tuesday. Hence, any state-specific information the voters receive will move vote shares more in the former than in the latter set of states.[18]

The Democratic primaries have the nice feature that candidate withdrawals are rather well defined events. Importantly, for the purpose of fitting the data to our model, weak candidates do not "stay in the race" beyond the point that voters abandon them because they are perceived to have no chance. In contrast, the 2008 and 2012 Republican primaries are not as clean in this respect. Huckabee stayed past the point of when his candidacy became quixotic, and for this reason we terminate the Republican primary after the 34th state. Similarly, Republican candidates were slow to concede in 2012, so we cut off the race after 30 states (beyond which Romney was pulling around 5 times the votes of his nearest rival).[19]

3.2 Identification

Our data consists of the number of candidates competing in each state, their political position and vote shares, and the timing of state contests. We do not observe voter signals, the distribution of voters to political positions (μ^s), or the candidate valence. With data from only three primary runs, respectively, it is not feasible to obtain credible estimates of μ^s; we instead posit that μ^s is a random draw from the uniform distribution with mean equal to one half and support S_μ.[20] Moreover, since most candidates compete in a small number of contests, candidate specific valence estimates would not be credible. Given that we do not estimate state specific values of μ^s, inverting the vote shares to obtain the state signals is not feasible.[21] Rather, we only aim to estimate (i) the standard deviation of candidate valence, σ_v; (ii) the standard deviation of state-specific information shocks, σ_η; (iii) the salience of the two major political positions, λ; and (iv) the support of electoral preferences for the two main political positions, S_μ. These parameters are sufficient to estimate the distribution of vote shares for a set of candidates competing in a state.[22]

We now turn to an informal discussion of identification. To provide some intuition about the main sources of identification, we consider subsets of the four parameters separately, even though all four parameters are estimated jointly and more than one source of variation in the data helps to pin down any given parameter.

The parameters S_μ and σ_η are identified jointly from the level and time variation of vote share volatility. Holding the candidates fixed, the model predicts that vote share volatility declines over time as voter beliefs about candidates'

[18] The realized vote share volatility in the first 5 states is very noisy because it is based on very few states (in contrast, there are more than 20 states in each of the latter two groups of states). Moreover, in the first five states, three instead of two candidates compete, which tends to decrease each candidate's vote share volatility there. For this reason, it is theoretically ambiguous whether we should expect vote share volatility in the first five states to be higher or lower than on Super Tuesday.

[19] Missouri is also dropped because Gingrich was not on the ballot for a technicality.

[20] Deltas and Polborn (2015) find that the political positions of the candidates (i.e., "insider" versus "outsider" or "conservative" versus "moderate") do not significantly affect the candidates vote shares in the Presidential primaries. This finding can be used as a (rough) justification for our assumption here that $E(\mu^s) = 1/2$.

[21] Inverting the vote share to obtain the signal would still contain the unknown value of μ^s on the right hand side.

[22] We consider candidate withdrawal as exogenous conditional on prior vote shares, i.e., we do not use it to draw any inference about state-specific signals for candidates beyond the withdrawal date. This appears reasonable: Most candidates stay in the race until lack of funding forces them to withdraw, and donor support is probably based on election performance and the corresponding public information, but not on a candidate's private information.

valence become more concentrated around the true value. In the limit, once candidates' valences become known, share variability is driven solely by the variability in μ^s. Thus, holding other parameters constant, S_μ is identified from the limit share variability, and σ_η is obtained from the rate of decline in share variability towards that limit. Of course, with finite primary runs, the limit is never reached. However, the rate of decline is informative of where the lower bound in vote share variability lies.

The parameters λ and σ_v are identified jointly, from the mean vote shares of candidates taken over groups of state contests, from the change in these vote shares after candidates withdraw, and from the average value of the maximum vote share over groups of state contests (again over groups of state contests). Holding other parameters constant, high values of λ imply that a higher percentage of voters whose first choice is a withdrawing candidate will vote for another candidate with the same political position as the withdrawing candidate (because candidates in the opposite political position are poor substitutes for the withdrawing candidate). The value of σ_v is identified from the share of the political positions as a function of the number of candidates in each position, both initially and in later election rounds, and from the average value of the maximum vote share. The higher the value of σ_v, the higher the expected difference in valence between the best candidate and the other candidates, and also the higher the expected difference in valence between the best of the candidates in a position with many candidates and the position with fewer candidates. Thus, higher values of σ_v are associated with higher maximum vote shares and also with lower vote shares for the political position with fewer candidates.

As noted above, identification of any particular parameter comes from multiple sources of data variation, and the informal discussion above focuses on the main sources of identification. To see the interdependence of parameter estimates, consider the following example with just one candidate in each position. In that case, a higher value of λ would increase the value of σ_η (or of S_μ) implied by any given observed vote share volatility of that candidate. Since that candidate would be a poorer substitute for the other candidates, higher vote share variability could be rationalized by higher signal volatility and/or higher variability of voter preferences across states. Similarly, changes in the two parameters that drive vote share volatility also have an impact on average shares (given that the vote share functions are non-linear). Our estimation procedure jointly pins down the parameter values from all these variations in the data, i.e., we do not match specific parameters to a specific moment in the data, but rather employ a standard generalized methods of moments framework.

Finally, note that the aggregate vote shares of the two political positions are sufficient for all of the above identification arguments to go through. Distinguishing the vote shares of individual candidates adds some information, but also substantial complexity, largely because we would have to estimate candidate specific valences in order to utilize candidate level data. With only a few observations for many of the candidates, any candidate specific estimates would be based on very limited information and be very imprecise.

3.3 Estimation

We estimate the unknown parameters S_μ, σ_v, σ_η, and λ using the Generalized Method of Moments, allowing the parameters to differ for the Democratic and Republican primaries. We employ 15 moments in total for the Democrats and 9 moments for the Republicans, and assume that all parameters are common to all elections within a party.

Note that, for the distribution parameters σ_v, σ_η and S_μ, this only assumes that the *distributions* from which the actual candidate realizations, state preferences and signals are drawn are equal in an ex-ante sense, but the actual realizations of the random variables can differ across elections. Assuming constancy of parameters is extremely

helpful for identification,[23] and appears plausible, given that the three elections are all within twelve years without a major realignment of the internal divisions in the parties, or fundamental changes in the way in which voters acquire information about candidates. Over the long run (say, over 20 years), though, it might be more of a stretch to assume that all parameters remain constant, and so extending the set of nomination races used in the estimation would not necessarily lead to "better" estimates of the relevant model parameters.

We now describe our estimation approach, starting with the Democratic primaries. Let W_y^s denote the observed vote share of all insider candidates in state s and year y.[24] For each year and state, we calculate the expectation of W_y^s, as given by the model, by integrating over the distribution of valence draws for the candidates competing in that year, the distribution of signal histories observed by the voters of state s for each of the candidates, and the distribution of μ^s. For each year and political position, we draw a number of valence values equal to the number of candidates in that political position for that year. For the early state contests, when all candidates are in the race, all valence draws are used for the vote share calculations. In later rounds, as candidates withdraw, we sequentially drop the valence that corresponds to the lowest *perceived* valence among candidates in a position when a candidate in that position withdraws (see Appendix for details).

We partition state contests into six groups, indexed G1 to G6, distinguished by the number of candidates in each political position and the contest year.[25] These groups are used in constructing the moment conditions, to which we turn next.

The first moment condition is based on the expected aggregate vote shares of the insider candidates, and provides identifying power primarily for the dispersion of valences, σ_v and the substitutability between candidates, λ. It is given by

$$E_{y,s}\left\{E_{v,\vec{\mu},\vec{\eta}}\left[W(v_y^s, \eta_s, \vec{\eta}_{\rho<\rho_s}, \mu_y^s, \sigma_\eta, \lambda, \sigma_v)|y, s\right] - W_y^s\right\} = 0, \tag{6}$$

where $W(\cdot)$ is the model's predicted vote share of the insider candidates as a function of valences, signal noise draws, signal noise histories, and state preferences and other parameter values, vector v_y^s is the set of valence values of year y for the candidates who are competing in state s, as outlined earlier, while $\vec{\eta}_{\rho<\rho_s}$ is the sequence of signal noise draws for these candidates for the rounds of voting that precede state s. The outer expectation that forms this, and all the remaining, moment condition is taken over all three nomination contests. The (inside) expectation of the predicted vote share in state s in year y is taken with respect to the distribution of sequences of μ_y^s, $\vec{\mu}$, the signal noise sequences for all candidates, $\vec{\eta}$, and valence value vectors. This expectation conditions on the number of competing candidates and the full history. The pair of year and state is sufficient information for obtaining the expectation, which is why the expectation is explicitly conditioned on the pair (y, s), even though such conditioning is implicit in the arguments of $W(\cdot)$. Thus, the difference between $E_{v,\vec{\mu},\vec{\eta}}[W(\cdot|y, s)]$ and W_y^s is uncorrelated with the number of competing candidates, the round in which the state votes, or any other variables that are fully determined by the year, y, and round ρ. Indicator variables for each of the six state contest groups are a function of the current number of candidates competing in a state and the initial number of candidates competing in the primary, and thus satisfy this requirement.[26] Let g denote

[23] It is technically possible to allow one parameter to differ across years, but even with keeping the other three constant throughout all years, identification will be much weaker.

[24] The estimation results are completely invariant as to which of the two political positions we choose.

[25] The groups are as follows G1: 2000, all rounds (Gore, Bradley); G2: 2004, rounds 2 and 3 (Kerry, Edwards, Dean, Clark, Lieberman); G3: 2004, rounds 4 to 7 (Kerry, Edwards, Dean; and Clark for rounds 4 and 5); G4: 2004, rounds 8 and 9 (Kerry, Edwards); G5: 2008, rounds 1 to 5 (Obama, Clinton, Edwards); G6: 2008, rounds 6 to 17 (Obama, Clinton).

[26] G2 is a composite group, because there is only a very small number of states in which Dean is running but Clark is not. This composition does not create any issues with regards to the estimation procedure.

a state contest group. We then obtain the following moments:

$$E_{y,s}\left\{\left(E_{v,\vec{\mu},\vec{\eta}}\left[W(v_y^s,\eta_s,\vec{\eta}_{\rho<\rho_s},\mu_y^s,\sigma_\eta,\lambda,\sigma_v)|y,s\right]-W_y^s\right)\mathbf{1}_{(s,y)\in g}\right\}=0 \tag{7}$$

where $\mathbf{1}_{(s,y)\in g}$ takes the value of 1 if state s voted in year y in the contest group g, and zero otherwise (observe that g is a partition of state-year pairs). This exhaustive interaction with the dummies makes inclusion of moment condition (6) redundant, and thus we drop it. Notice that for state contest group G1 (Bradley vs. Gore), the expectation of $W(\cdot|y,s)$ is equal to $\frac{1}{2}$ for all parameter values. Thus, this condition has no identifying power and is also dropped. Therefore, g takes the values of {G2, G3, G4, G5, G6} yielding a total of five moments. Details about the computation of all moments are given in the Appendix.

The second set of moment conditions is based on the expected maximum vote share in the final part of each primary race, when only two candidates remain. This condition provides strong identifying power for σ_v, since the higher the value of σ_v the higher the expected maximum vote share, and the more responsive it is to the number of candidates initially participating in the race. The moment condition also provides auxiliary identifying power for λ. Analogous to the construction of (7), this condition results in a moment for each of the three nomination races, formally written as

$$E_{y,s}\left\{\left(E_{v,\vec{\mu},\vec{\eta}}\left[\max\{W(\cdot),1-W(\cdot)\}|y,s\right]-\max\{W_y^s,1-W_y^s\}\right)\mathbf{1}_{(s,y)\in g}\right\}=0, \tag{8}$$

where the arguments for $W(\cdot)$ are the same as in (7), and $g\in\{G1,G4,G6\}$, yielding a total of three moments.

The moments described to this point are all first moments. We also employ second moments based on the share variability of $W(\cdot|y,s)$. These moments provide information relevant to pinning down S_μ and σ_η. Following the above construct, they are given by

$$E_{y,s}\left\{\left(E_{v,\vec{\mu},\vec{\eta}}\left[\left|W(v_y^s,\eta_s,\vec{\eta}_{\rho<\rho_s},\mu_y^s,\sigma_\eta,\lambda,\sigma_v)-\tilde{W}_{g_{y,s}}\right|\mid y,s\right]-\left|W_y^s-\bar{W}_{g_{y,s}}\right|\right)\mathbf{1}_{(s,y)\in g}\right\}=0, \tag{9}$$

where the random variable $\tilde{W}_{g_{y,s}}$ is the mean aggregate vote share of the insider candidates for a particular sequence of signal noise and state voter preferences in the corresponding state contest group, and $\bar{W}_{g_{y,s}}$ is the average vote share of these candidates in the same states in the data. Specifically, the mean vote share in any contest group g is defined as $\tilde{W}_g=\frac{1}{\|g\|}\sum_{(s,y)\in g}W(v_y^s,\eta_s,\vec{\eta}_{\rho<\rho_s},\mu_y^s,\sigma_\eta,\lambda,\sigma_v)$, and the sample analogue is defined as $\bar{W}_g=\frac{1}{\|g\|}\sum_{(s,y)\in g}W_y^s$, where $\|g\|$ indicates the number of states in g. Notice that $W(\cdot)$ does not have the same expected value for every state in g because states in g differ in the number of prior signals (in the case of G3, they also differ in the number of candidates). Therefore, the expectation of the absolute value in (9) differs when evaluated in sub-groups of g. Subdividing the state contest groups so that each group includes only states that vote in the same round would result in groups that are too small for a reasonable estimate of the sample mean $\bar{W}_g$. We thus include contest groups G1, G4, and G6 which consist of at least 10 states, even if some of these states did not vote in the same round. We index these state contest groups by $g2$. Incidentally, these groups consist of contests between only two candidates.[27] Note the moment condition (9) can be satisfied by systematically overshooting the difference between the predicted and the observed absolute deviations from the mean for states voting early within a group and undershooting it for states that vote in later rounds (or vice versa). The time profile of differences between predicted and observed share variability is not relevant. It is in this sense that these moments provide stronger identification power for S_μ than for σ_η.

The next two moment conditions rectify this overshooting/undershooting issue, and are designed to help pin down σ_η, i.e., the speed of learning about candidate valence. They are calculated on state contest groups that are sufficiently

[27] In principle, we could compute this moment condition over all states, but most identification power would be lost since the mean vote share strongly depends on the number of candidates.

long for meaningful changes in vote share variability. The first one is the Bradley versus Gore contest, which took place over 15 states (group G1). We subdivide these elections into group G1a, consisting of the first seven (which took place over 5 rounds), and group G1b, consisting of the remaining eight (which took place in a single round). The second case is the Obama versus Clinton contest which took place over 45 states (group G6). We subdivide these elections into group G6a, consisting of the 22 Super Tuesday elections, and group G6b, consisting of the remaining 23 elections (over 9 rounds).

Define the variable $\mathbf{D1}_y^s$ to take the value of 1 if $(s, y) \in G1a$, -1 if $(s, y) \in G1b$, and 0 otherwise. Similarly, let $\mathbf{D6}_y^s$ take the value of 1 if $(s, y) \in G6a$, -1 if $(s, y) \in G6b$, and 0 otherwise. The two moments conditions are then given by

$$E_{y,s}\left\{\left(E_{v,\vec{\mu},\vec{\eta}}\left[\left|W(v_y^s, \eta_s, \vec{\eta}_{\rho<\rho_s}, \mu_y^s, \sigma_\eta, \lambda, \sigma_v) - \tilde{W}_{g_{y,s}}\right| \mid y, s\right] - \left|W_y^s - \bar{W}_{g_{y,s}}\right|\right)\mathbf{D1}_y^s\right\} = 0 \tag{10}$$

and

$$E_{y,s}\left\{\left(E_{v,\vec{\mu},\vec{\eta}}\left[\left|W(v_y^s, \eta_s, \vec{\eta}_{\rho<\rho_s}, \mu_y^s, \sigma_\eta, \lambda, \sigma_v) - \tilde{W}_{g_{y,s}}\right| \mid y, s\right] - \left|W_y^s - \bar{W}_{g_{y,s}}\right|\right)\mathbf{D6}_y^s\right\} = 0. \tag{11}$$

A few things are worth pointing out in these moment conditions. First, deviations from the predicted and observed absolute deviation in the two sub-groups are weighted with opposite signs, and thus cannot cancel out (unless they are of the same sign, in which case this would lead to violations of (9)). Second, sample standard deviation that are the same in the early and late sub-groups of G1 and G6 imply no further updating of valences, and thus near zero values of σ_η. By contrast, large reductions in the sample standard deviation require higher values of σ_η to ensure that the moment conditions are satisfied. Third, the value of $\tilde{W}_{G1}$ is a weighted average of $\tilde{W}_{G1a}$ and $\tilde{W}_{G1b}$, which are not in general equal to each other (and similarly for G6a and G6b and for their sample analogues).[28]

The final group of moments is based on the expected change in the insider candidate's vote shares after some of the candidates in the outsider position withdraw. These moments provide strong identification of the parameter λ, as they measure the gain of a candidate when a competitor in the other political position withdraws, for a given set of valence draws (of course, this is then integrated over the distribution of valences). There is no exit in the 2000 primary, so these moments are operative only for the 2004 primary, where they measure the percentage increase in Kerry's vote share after Dean and Clark withdraw, and in the 2008 primary where they measure the percentage increase in Clinton's vote share after Edwards withdraws. Following the previous constructs, these moments are given by

$$E_{y,s}\left\{\left(E_{v,\vec{\mu},\vec{\eta}}\left[\left.\frac{W(v_y^s, \eta_s, \vec{\eta}_{\rho<\rho_s}, \mu_y^s, \sigma_\eta, \lambda, \sigma_v)}{\tilde{W}_{gLag_{y,s}}}\right| y, s\right] - \frac{W_y^s}{\bar{W}_{gLag_{y,s}}}\right)\mathbf{1}_{(s,y)\in g}\right\} = 0, \tag{12}$$

where $gLag_{y,s}$ is the group of states preceding the group that the state-year combination (y, s) belongs to, and g takes the values $\{G4, G6\}$. This completes the description of the 15 moment conditions for the Democratic primaries. With the number of parameters being fewer than the number of moment conditions, there exist no parameter values that would satisfy all of the equations; thus we obtain parameter estimates using GMM. Further technical details are in the Appendix.

The Republican primaries employ the same type of moments, using the vote share of the "moderate" candidates. The rounds in the Republican primaries are divided into five groups (some sub-divided to subgroups). Γ1: 2000, all rounds, consisting of 21 states (Bush, McCain); Γ2: 2008, rounds 1 to 5, with 6 states (Guiliani, Huckabee, McCain, Romney, Thomson); Γ3: 2008, rounds 6 and 7 with 21 states (Huckabee, McCain, Romney), Γ4: 2008, rounds 8 to 10

[28] One could construct more complicated moment conditions where the absolute value of the differences are taken from those sub-groups averages. We chose not to do so since the sample vote share averages do not vary much between the early and the late sub-groups for either primary race, and the mean of the sub-groups would have been based on a very small number of states.

with 7 states (Huckabee, McCain), and Γ5: 2012, all rounds, consisting of 29 states (Gingrich, Romney, Santorum). The estimation methodology is otherwise the same. The first six rounds of Γ1 form subgroup Γ1a while the last round, consisting of 12 states, forms subgroup Γ1b. For moment condition (7), g takes the value of Γ3, Γ4, and Γ5, while for moment condition (8), g takes the value Γ1, Γ4, and Γ5.[29] The vote share variability moment (condition (9)) is estimated for Γ1, Γ3 and Γ5, while the change in share variability moment is estimated in a fashion analogous to (10) and (11) using the subgroups Γ1a and Γ1b. The last moment, which is analogous to that of (12) is estimated based on the change in McCain's vote share after Romney's withdrawal, comparing the relative shares of McCain in Γ3 and Γ4.

3.4 Estimation Results

The estimation results and associated (asymptotic) standard errors are $\hat{\sigma}_v = 0.59 \pm 0.73$, $\hat{\sigma}_\eta = 0.47 \pm 6.4$, $\hat{\lambda} = 1.9 \pm 1.16$, and $S_\mu = 0.53 \pm 0.19$ for the Democrats, and $\hat{\sigma}_v = 1.57 \pm 4.20$, $\hat{\sigma}_\eta = 3.05 \pm 28.2$, $\hat{\lambda} = 3.00 \pm 2.41$, and $S_\mu = 0.56 \pm 0.33$ for the Republicans. The standard errors are large in part because a common set of parameters are used to fit three distinct primaries, respectively. Thus, parameters try to match some "average" primary, resulting in somewhat poor fit to all of them and higher standard errors. Nonetheless, the standard errors are indicative of the relative confidence in our point estimates, with the dispersion in voter preferences being most precisely estimated and the variance of signals being least precisely estimated. In our simulations, which is where the main interest of our paper lies, we carry extensive sensitivity analysis to ensure that our findings are robust to parameters that differ substantially from the point estimates. However, before we proceed to these simulations, it is useful to briefly discuss the relative importance of candidate valence, voter preferences, differences in these preferences across states, and voter uncertainty about candidates implied by the point estimates.

The Democratic point estimate of σ_v indicates that the better of two candidates in the same political position who differ in one standard deviation of valence will obtain $\Phi(0.59) \approx 72\%$ of the voters who share the same political position when voters knew the true valences perfectly. (Remember that the standard deviation of idiosyncratic preference shocks, σ_ε, is normalized to 1, so that Φ is the cdf of ε.) Among Republicans, this value is $\Phi(1.57) \approx 94\%$.

The point estimate of λ indicates that a candidate in position 0 who is one standard deviation better (in terms of valence) than a candidate in position 1, among Democrats, will obtain $\Phi(2.49) \approx 99\%$ of the voters in position 0 and $\Phi(-1.31) \approx 10\%$ of the voters in position 1. Two candidates of equal valence but different positions get $\Phi(1.9) \approx 97\%$ of the voters with the same position and $\Phi(-1.9) \approx 3\%$ of the voters with the opposite position. Among Republicans, a candidate in position 0 who is one standard deviation better than a candidate in position 1 will obtain $\Phi(4.57) \approx 100\%$ of the voters in position 0 and $\Phi(-1.43) \approx 8\%$ of the voters in position 1. Two candidates of equal valence but different positions get $\Phi(3.00) \approx 99.86\%$ of the voters with the same position and $\Phi(-3.00) \approx 0.14\%$ of the voters with the opposite position. Thus, the estimation results imply that political positions are very important for voters in both parties, but particularly among Republicans.

The point estimate of σ_η indicates that uncertainty about candidate valence is substantial in the states that vote early. Consider Candidate j's perceived valence after N signals have been observed, $\hat{v}_j^N$. From an ex-ante point of view (i.e., before valence and signal realizations have been drawn), this is a random variable with expected value 0 (by the fact that the expected value of valence is zero, and expectations after signals follow a martingale). Given realized

[29] The 2000 primary has only two (serious) candidates, and thus moment (7) has no identifying power.

signals $(Z_j^s)_{s=1..N}$, expected valence is[30]

$$\frac{\sigma_v^2}{\sigma_v^2 + \frac{\sigma_\eta^2}{N}} \cdot \frac{\sum_{s=1}^N Z_j^s}{N}.$$

Thus, the variance of perceived valence after N signals have been observed is

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left(\frac{\sigma_v^2}{\sigma_v^2 + \frac{\sigma_\eta^2}{N}} x \right)^2 \phi\left(\frac{x - v}{\sqrt{\sigma_\eta^2/N}} \right) dx \quad \phi(v/\sigma_v) dv = \frac{\sigma_v^4}{\sigma_v^2 + \frac{\sigma_\eta^2}{N}}. \tag{13}$$

Note that this variance is always smaller than σ_v^2, because signal uncertainty implies that non-mean realizations of v are only learned over time, and the fact that voters know that signals are imperfect means that their updating process is dampened. Moreover, the variance of perceived valence is increasing in N and goes to σ_v^2 in the limit of $N \to \infty$; this is intuitive because, when valence is eventually revealed, the variance of perceived valence is the same as the ex-ante variance of valence. For the Democratic point estimates, (13) implies that the standard deviation of perceived valence is approximately 0.46 in the first district, about 0.56 by the fifth district, and about 0.58 for district 25. For the Republican parameters, the standard deviation of perceived valence is approximately 0.72 in the first district, about 1.19 by the fifth district, and about 1.46 for district 25. Thus, the perceived valence difference between the two candidates is initially (in expectation) substantially smaller than the true valence difference, so that there is substantial vote-splitting between two candidates in the same position.

Finally, the point estimate of the support of μ indicates that the percentage of voters in each political position varies between about 1/4 and 3/4 for each party. In the typical state, in terms of deviation from the 50/50 voter partition, about five-eighths of the voters support one position and three-eighths support the opposite. However, if the candidates have equal perceived valence, then vote shares are less variable than μ because candidates obtain symmetric vote shares from voters in both positions, and the candidate in the less popular position benefits more from relatively frequent cross-over voters with majority preferences than his competitor benefits from cross-over voter with minority preferences.[31]

4 Simulated effects of different institutions

We now use the point estimates of the parameters to generate a baseline scenario quantifying the performance of different primary systems. In Section 4.2, we analyze the robustness of these results to parameter changes.

4.1 The baseline scenario

Our basic approach is as follows: We always consider races with three candidates, two of whom share a position while the third one is in the other position. In each simulation run for the Democrats, we first draw candidate valences from the estimated normal distribution $N(0, 0.59)$. Among the candidates who share a position, this generates two

[30] This is a weighted average of the ex-ante expected valence, 0, and the average signal realization (the second fraction), where the weight depends on the precisions of the ex-ante distribution of v and the precision of the signal distribution for N signals.

[31] For example, among Democrats, in the typical state in terms of deviation from a 50/50 voter partition, the candidate with the less popular position in the state obtains $\frac{3}{8}97 + \frac{5}{8}3 \approx 38.25$ percent of the votes and the candidate with the more popular position obtains 61.75 percent of the votes.

candidates with different valences, whom we denote B (for "better") and W (for "worse"). The other, "solitary", candidate is denoted by S.

We then draw state-specific signals according to $N(0, 0.47)$. Depending on the temporal structure of elections (and hence, on which signals are effectively observable in a state), this generates, according to Bayesian updating, voters' beliefs in a state. We also draw aggregate position preferences in state s, μ^s, from a uniform distribution on $[0.235, 0.765]$. Together with the distribution of individual preference shocks, drawn from $N(0, 1)$, this generates the vote distribution for candidates in a state. Aggregating over all states, we find the average vote share of each candidate, and the candidate with the most votes wins the nomination for a given run. (In the simulation, all states have the same size so that a candidate's aggregate vote share is simply the unweighted average of the candidate's vote shares in all states). We repeat this process 80,000 times to generate a probability distribution over outcomes, e.g., the proportion of times that B, W and S win the nomination. Our procedure for the Republicans is analogous, but replacing the parameter values estimated in the Democratic primaries by those estimated from the Republican ones.

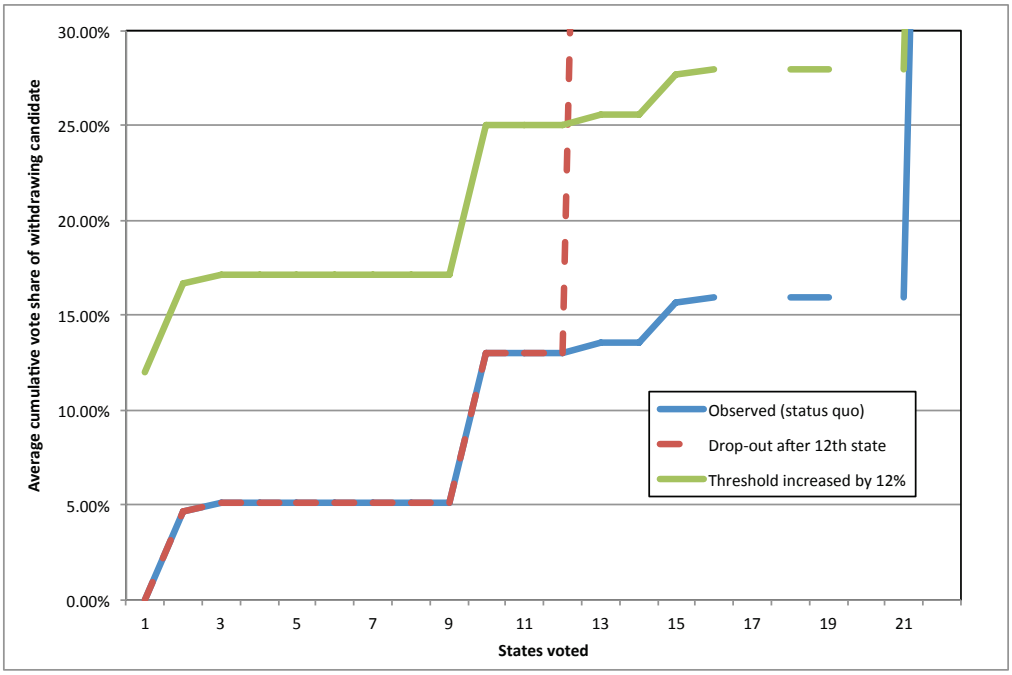

Figure 2: Exit threshold function

We model exit decisions as follows (see Figure 2). We define an *exit threshold function* such that, if a candidate falls below the exit threshold in some period/state then we withdraw him for all remaining rounds. The exit threshold function that we use is constructed from the empirical observations of the 2000–2008 Democratic primaries as follows: For each voting round up to state 16, the exit threshold is the lowest cumulative vote share observed from any active – i.e., not yet exited – candidate (among those used in the estimation). In all three nomination contests, the field was narrowed down to 2 candidates by state 16, so that there are no observations for any of the subsequent states. We therefore have to choose the value of the exit threshold function after state 16 in a somewhat more arbitrary way. We do this by setting the exit threshold for the next 5 states to the same value as in state 16 and then increase the threshold to 1 at state 21 (i.e., we assume that the third candidate has to exit after state 21).

For the Republicans, modeling exit decisions based on observed withdrawals would be somewhat problematic, both because of a limited number of observations and because of some idiosyncratic withdrawal decisions; for example, in 2008, Romney withdrew before Huckabee, even though he had a considerably higher aggregate vote share at the time. Furthermore, many observers at the time believed that Huckabee stayed in the race in order to make sure that Romney would not win, a motivation not in line with a simple tradeoff calculation between the cost of staying in and the chance of winning. Consequently, we do not estimate a separate exit function for Republicans, but rather use the same function as for Democrats. In interpreting the results, however, we note that Republicans tend to stay longer in the race than the Democrats.

We compare three basic primary systems. The first system is a completely sequential primary in which only one state votes at any given time. All three candidates compete until the candidate from the two that share a common position drops out (as his vote share drops below the exit threshold), while the remaining two candidates compete in the remaining districts. The second system is a completely simultaneous primary in which all states vote at the same time. The third system follows the NASS proposal, i.e., there are initially two states that vote sequentially, followed by 4 groups of 12 states, respectively, that each vote simultaneously. Candidate withdrawal approximates the empirical drop-out rule; the weakest candidate among B and W withdraws before the vote in the second large group of states on the basis of the vote shares in the preceding group of states, unless he has not met the drop-out threshold in the first two states, in which case he withdraws earlier.

The last two columns represent variations on the sequential system. Specifically, in the fourth column, we assume that the third candidate, the one who is worse perceived among the two sharing the same position, has to drop out in state 9 among Democrats and state 15 among Republicans, while in the fifth column, we shift up the exit threshold function by 13 percentage points for Democrats and by 1 percentage point for Republicans. Both of these changes lead to earlier dropout relative to the baseline system I.[32] These changes to the dropout rules are chosen such that the rule is (approximately) optimal for the two parties, respectively. See our discussion below for more details. Tables 1 and 2 summarize the results for the Democratic and Republican parameters, respectively.

The first and second three rows in both tables provide the mean vote shares and winning percentages of candidates S, B and W in the different primary systems, respectively. The next three rows report the average valence in the different primary systems, conditional on S, B or W winning the nomination. The next two rows give the winning probabilities of candidates S and B, conditional on being the full information Condorcet winner. (Remember that Candidate W is never the Condorcet winner, because his position is the same as that of Candidate B, and his valence is lower).[33] Finally, the last two rows report the overall probability that the Condorcet winner wins, and the winner's expected valence.

[32] Since candidates exit when their cumulative vote share falls below the threshold for the first time, an increase in the dropout rule corresponds to earlier withdrawal of candidates. Similarly, system IV forces all third candidates to drop out at the given state, even those who would continue to stick around for a few more elections in system I.

[33] Hence, all voters with $\varepsilon_W \leq \varepsilon_B$ (i.e., half of the population) strictly prefer B over W. By continuity, the set of voters who prefer B to W is always larger than the set of voters who prefer W to B.

	I: Sequential	II: Simultaneous	III: NASS	IV: Sequential w/ fixed dropout (state 9)	V: Sequential w/ higher threshold (+13%)
S vote share	44.5%	42.6%	45.1%	45.3%	45.5%
B vote share	46.0%	35.2%	45.1%	47.7%	47.5%
W vote share	9.6%	22.1%	9.9%	6.9%	7.02%
S wins	43.6%	76.7%	41.3%	38.2%	37.8%
B wins	54.0%	23.1%	54.1%	58.3%	58.8%
W wins	2.5%	0.2%	4.6%	3.5%	3.5%
E[val./S wins]	0.382	0.150	0.405	0.436	0.432
. . . B wins	0.562	0.783	0.553	0.534	0.533
. . . W wins	0.193	0.432	0.136	0.144	0.148
S wins if CW	98.7%	100%	97.7%	97.1%	95.9%
B wins if CW	80.1%	34.6%	79.9%	85.8%	86.0%
Prob[CW wins]	86.3%	56.3%	85.8%	89.6%	89.2%
Winner's E[val.]	0.475	0.297	0.472	0.483	0.482

Table 1: Simulation results for Democratic parameters

	I: Sequential	II: Simultaneous	III: NASS	IV: Sequential w/ fixed dropout (state 15)	V: Sequential w/ higher threshold (+1 %)
S vote share	43.9%	46.7%	45.0%	43.9%	44.1%
B vote share	46.0%	31.4%	41.5%	46.3%	45.6%
W vote share	10.1%	21.8%	13.5%	9.9%	10.4%
S wins	43.2%	98.3%	44.7%	42.9%	41.9%
B wins	51.1%	1.7%	46.6%	51.5%	51.5%
W wins	5.7%	<0.1%	8.7%	5.6%	6.6%
E[val.|S wins]	0.941	0.0287	0.9052	0.959	0.976
. . . B wins	1.466	3.1626	1.469	1.465	1.451
. . . W wins	0.259	2.1138	0.139	0.257	0.217
S wins if CW	90.7%	100%	90.4%	91.2%	90.1%
B wins if CW	72.3%	2.6%	65.5%	73.1%	72.6%
Prob[CW wins]	78.4%	34.9%	73.8%	79.1%	78.4%
Winner's E[val.]	1.170	0.0824	1.101	1.180	1.170

Table 2: Simulation results for Republican parameters

Democratic parameters (Table 1) Comparing the first three systems for the Democratic parameters shows that, from a welfare perspective, the sequential benchmark (regime I) performs marginally better than the NASS Proposal (regime III), independent of whether this performance is measured by the probability that the Condorcet winner wins, or the winner's expected valence. Simultaneous voting in all 50 states (regime II) does worst by a significant margin.

As for the intuition, consider first the simultaneous system in which candidates B and W split the votes among themselves in all districts. As a consequence, Candidate S wins almost 77% of all races, including $100 - 35 = 65\%$ of those races in which candidate B would have been the Condorcet winner (Candidate S also always wins if he is the Condorcet winner). The expected valence of the election winner is thus quite close to zero, the ex-ante expected valence of Candidate S. Candidate B has a chance of winning only when he is significantly better than both Candidate S and Candidate W. Therefore, B's valence in those few instances where he wins is actually very high (about 1.3 standard deviations above the expected valence).

Now consider regime I, the purely sequential system. The learning facilitated by the sequential structure has the effect that votes shift from W to B, while S's vote share is just a bit higher than in regime II.[34] Consequently, B now wins much more often (54% of races). Note, however, that Candidate S still has an advantage in this system, as S continues to win in many cases when he is not the Condorcet winner. This is reflected in the candidates' winning probability conditional on being Condorcet winner: While S wins 98.7% of the races when he is the Condorcet winner, B wins only with probability 80.1% when he is the Condorcet winner.[35]

The quantitative difference between the simultaneous and sequential system is substantial in terms of the welfare measures: The probability that the Condorcet winner is selected as nominee increases from 56.3% under simultaneous voting to 86.3% under sequential voting. The expected valence increase of $0.475 - 0.297 = 0.178$ is equal to approximately $0.3\sigma_v$.

As for regime III, our simulation shows that the partially sequential NASS system does only marginally worse than the fully sequential system: The probability that the election winner is the Condorcet winner and the expected valence of the winner are slightly smaller in regime III than in regime I, but substantially larger than for the simultaneous system. The main difference between regimes III and I is due to the fact that, with vote splitting likely to occur at least for one large round (i.e., $2 + 12 = 14$ states), the probability that S wins when he is the Condorcet winner decreases by 1 percentage point.

Republican parameters (Table 2). Recall that the Republican parameter values are quite different from the Democratic ones: The importance of the different positions, λ, is considerably higher among Republicans, as is the uncertainty about candidate quality. In spite of these differences, the results for the Republican parameters are qualitatively very similar — specifically, simultaneous voting again does significantly worse than the different sequential systems,

[34] Whenever a system induces the exit of a candidate, any remaining candidate's vote share must increase, since, because of the idiosyncratic shock, there are always voters who rank the exiting candidate first and that particular remaining candidate second.

[35] The reason that B wins absolutely more often than S is that B's expected valence is higher than S's, since he is the better of two candidates in his position – as valence draws are iid, the probability that B's valence is higher than S's is $2/3$.

and the partially sequential NASS system III does marginally worse than the baseline sequential system I. In comparison to the Democratic parameters, the difference in the performance of systems I and III is a bit bigger here. Intuitively, there is more miscoordination in system III, and this disadvantage becomes more important because of the increase in λ among Republicans.

Changes to dropout rules. An interesting question in the sequential system is whether candidates drop out too early or too late relative to what would be socially optimal in order to select the best candidate. Note that, in principle, the party can affect dropout decisions at the margin, for example by providing a number of superdelegate slots to each of the top two candidates after a pre-specified round, which would make it harder for the third candidate to win, and thus encourages him to drop out. In contrast, if the party wants to keep candidates in the race for longer, they could provide additional resources uniformly to all candidates, which makes it easier for a lagging candidate to stick around without having to rely exclusively on private fundraising.

To analyze this question, we consider the exit threshold rule described above, and then generate new rules by requiring the third candidate to drop out by state 9 among Democrats and state 15 among Republicans (column 4), and by uniformly adding 13 percentage points among Democrats and 1 percentage point among Republicans to the empirical rule used for column 1 (column 5). These rule variations are chosen such that they are optimal in their respective classes for the Democratic and Republican parameters, respectively. For example, among all exit threshold rules that shift up or down the empirical exit function by a fixed percentage, the one that adds 13 percentage points is optimal for the Democratic parameters.

Systems IV and V increase the probability that the Condorcet winner wins by about 3 percentage points among Democrats, and by less than 1 percentage point among Republicans, relative to the baseline. Likewise, the winner's expected valence increases.

The optimal change in dropout rule is larger for Democrats than Republicans. Intuitively, since both σ_v and σ_η are larger in the Republican party, a longer period of learning about valence is desirable. Moreover, as explained above, peculiarities of the Republican primaries in 2008 and 2012 made it difficult to estimate a credible Republican exit function, so the reader should recall that the "empirical dropout function" underlying system I is derived from Democratic data only. If it is indeed the case that Republican candidates tend to stay in the race longer than Democratic candidates, then an optimal adjustment in the drop-out thresholds that leads to earlier dropout would yield greater benefits than those suggested by the above tables.

Comparison with the existing literature It is interesting to compare our results with Knight and Hummel (2015) who also analyze the welfare effects of sequential and simultaneous primaries. The crucial difference between our models is that all candidates in their model differ only in their valence, not their positions, which removes the advantage of sequential voting in our model, namely better coordination. With sufficiently many states and ex-ante symmetric candidates, a simultaneous primary system would guarantee in their model that the full information Condorcet winner wins the nomination. In contrast, a sequential system can lead to aggregation failure because early state signals receive

too much attention as they do not just influence the election in the state in which they are received, but all following states.[36]

However, Knight and Hummel (2015) allow for a different potential advantage of sequential elections. Candidate quality is drawn from different ex-ante distributions, so that sequential elections allow "dark horse candidates" – whose quality is drawn from a less favorable ex-ante distribution, but who might turn out more positive with some probability – to emerge more easily than they would in simultaneous elections. It turns out that this advantage, in their estimation, is insufficient to overcome the advantage of better statistical signal aggregation in simultaneous elections. In contrast, the reduction of vote-splitting in sequential elections in our model is sufficiently large to make them perform a lot better than simultaneous elections. Also, their model shows that our result that sequential elections perform better than simultaneous ones would likely be strengthened if we allowed for candidates to be drawn from different ex-ante distributions.

4.2 Robustness

4.2.1 Parameter changes

As argued above, the empirical analysis is necessarily based on data from only a few primaries, and even if it were to perfectly estimate the parameter values for these primaries, we would still worry that parameters are likely to change over time. Our interest in institutional design is to find primary structures that would do well for a number of different plausible parameter constellations. Therefore, we want to see how substantial changes in the parameters affect our main qualitative results. Because of the issues with the Republican contests explained above, we focus on the Democratic side as the baseline case in this section.

Specifically, we analyze how doubling or halving λ, σ_v and σ_η, respectively, while fixing the other three parameters at their level in the baseline case, affect results. For S_μ (the size of the support of the state preference realization μ), we analyze adding or subtracting 0.1 to the parameter estimate instead.[37] Table 3 provides the performance of the three systems from the previous section for the baseline and eight parameter changes.

Evidently, both welfare measures often change substantially when parameters change, but the qualitative effect on the ranking of the three systems is limited in that the sequential benchmark and the NASS system perform similarly well and both do generally considerably better than the simultaneous system.

The exception to this rule is the case of low λ where all systems perform similarly and simultaneous elections are even slightly better than the NASS proposal, with system I being best by a very small margin. For low λ, the three systems are almost equivalent due to the fact that the horizontal differentiation among candidates matters less and thus,

[36] For example, Knight and Schiff (2010) show that, if the sequence of states voting were rearranged randomly in the 2004 primary (but all state signals remain the same), then Edwards would have an 11 percent chance of winning the nomination.

[37] This reflects the lower uncertainty about the value of this parameter.

	I: Sequential	II: Simultaneous	III: NASS	IV: Sequential w/ fixed dropout (state 9)	V: Sequential w/ higher threshold (plus 13%)
Baseline case					
expected valence	0.475	0.297	0.472	0.483	0.482
CW wins	86.3%	56.3%	85.8%	89.6%	89.2%
$\lambda \uparrow$ ($\lambda = 3.8$)					
expected valence	0.173	0.003	0.186	0.239	0.266
CW wins	52.7%	35.8%	55.4%	64.4%	68.5%
$\lambda \downarrow$ ($\lambda = 0.95$)					
expected valence	0.492	0.486	0.484	0.491	0.491
CW wins	91.6%	88.9%	88.7%	91.2%	91.2%
$\sigma_v \uparrow$ ($\sigma_v = 1.18$)					
expected valence	0.990	0.965	0.988	0.992	0.992
CW wins	94.1%	88.5%	93.2%	94.9%	94.7%
$\sigma_v \downarrow$ ($\sigma_v = 0.295$)					
expected valence	0.188	0.000	0.194	0.214	0.211
CW wins	68.6%	33.5%	71.2%	78.9%	77.4%
$\sigma_\eta \uparrow$ ($\sigma_\eta = 0.94$)					
expected valence	0.453	0.036	0.436	0.459	0.457
CW wins	80.8%	35.1%	77.7%	82.4%	81.9%
$\sigma_\eta \downarrow$ ($\sigma_\eta = 0.235$)					
expected valence	0.481	0.395	0.485	0.490	0.490
CW wins	88.6%	69.1%	90.0%	92.9%	92.8%
$S_\mu \uparrow$ ($S_\mu = 0.63$)					
expected valence	0.473	0.297	0.470	0.481	0.479
CW wins	86.3%	56.7%	85.7%	89.4%	89.0%
$S_\mu \downarrow$ ($S_\mu = 0.43$)					
expected valence	0.476	0.296	0.474	0.485	0.484
CW wins	86.3%	56.0%	85.9%	89.7%	89.4%

Table 3: Results for different parameter values, relative to Democratic baseline

the three candidates become closer substitutes of each other. As λ decreases, the three candidates are on a more level playing field, as "vote splitting" occurs not only between B and W but also between S and the other two candidates. The opposite happens for high λ, which increases the potential vote splitting between B and W at the benefit of S, and hence absolutely deteriorates the performance of all systems, but relatively benefits systems I and III which allow for more coordination. The NASS proposal here dominates sequential voting slightly as it leads to faster coordination and thus a higher chance that B wins if he is the Condorcet winner.

As σ_v increases, the expected valence difference between candidates increases and is more likely to become decisive for voters' decisions. Thus, all systems become more likely to select the Condorcet winner as σ_v increases, and less likely to do so as σ_v decreases. Also, the winner's expected valence increases in σ_v because the winner is more likely to be the highest valence candidate, and the expected realization of the highest valence draw increases in σ_v. Intuitively, as $\sigma_v \to \infty$, all systems must deliver the same outcome, as almost always most voters agree on who is the best candidate. Considering the probability that the Condorcet winner wins the nomination as our measure of welfare, when σ_v increases, simultaneous elections reduce their disadvantage relative to the other two systems while the difference between the sequential benchmark and the NASS proposal remains pretty much unchanged.

In contrast, as σ_v decreases, valence becomes less important for voters, and votes between B and W will be split more equally as long as both are running. Thus, the electoral advantage of candidate S increases, so that he will almost always win in simultaneous primaries. Therefore, the expected valence of the winner is almost exactly equal to S's ex-ante expected valence of 0. In contrast, coordination allows for a substantial winning probability for one of the two candidates in the same position, and thus for a higher expected valence of the election winner in systems I and III, which perform essentially equally well.

If σ_η increases, then vote splitting becomes more severe because the quality of information about valence is smaller, so candidate S wins almost always under a simultaneous system. Thus, the probability that the Condorcet winner wins drops from around $1/2$ to around $1/3$ in that system. In contrast, while welfare is reduced in both sequential systems due to the less precise information, this effect is relatively much smaller, because they allow for coordination and therefore do not always lead to a victory of S. For high σ_η, the advantage of system I over system III is quite large because the NASS system forces relatively early coordination, and thus, leads to frequent eliminations of B. If σ_η decreases, then each signal is more informative about valence. For the two candidates in the same position, this effect diminishes the importance of vote splitting, which explains the improved relative performance of the simultaneous system for this case. Also, NASS proposal now slightly outperforms the sequential benchmark. Changes in S_μ have only a minimal quantitative effects relative to the baseline case, and thus, no qualitative changes occur.

Turning to columns IV and V confirms that the result that trailing candidates stick around too long from a social point of view appears qualitatively robust. Systems IV and V are always better than System I, except for the case of low λ (intuitively, when λ is small, the coordination issue is less severe, and so the disadvantage of earlier dropout rules – likelier mis-coordination on the wrong candidate – becomes relatively more costly).

4.2.2 Valence distribution for candidate S

In all simulations so far, we have assumed that the valence of candidates is drawn from the same distribution, independent of their position, and the number of candidates who compete in each position. Alternatively, the sole candidate S could be the result of some prior coordination among potential candidates with the same position. In this case, it is more reasonable to assume that the sole candidate's valence is drawn from a better distribution (because he is the winner of this internal competition).

A possible formalization of this idea is that there were two proto-candidates in position 0, but that, before the start of the primaries, the sole candidate already convinced the other candidate who was located in the same position, but had a lower valence, not to run. Thus, Candidate S's valence is $\max(v_{S,1}, v_{S,2})$, where $v_{s,j}$ is distributed $N(0, \sigma_v)$. As a consequence, the distribution of candidate S's valence is the same as the distribution of candidate B's valence, and each of them is the Condorcet winner with 50 percent probability.

In this case, the winner's expected valence in the sequential benchmark is 0.583, in the simultaneous primary is 0.459, and in the NASS system is 0.582, while the probability that the Condorcet winner wins the nomination is respectively 86.8%, 62.1%, and 87.0%. These results show that the performance of simultaneous elections in this scenario is substantially better than in the baseline case — essentially, because S now is the Condorcet winner more often and still wins with probability close to 1 —, while the effect in the two sequential systems is still positive but much smaller. In any case, the relative ranking of the three systems is again unaffected.

4.2.3 Changes in information quality in different institutional setups

All institutional comparisons presented so far are based on the assumption that the fundamental parameters remain the same across institutions. However, it is plausible that, if several states vote simultaneously in one same round, there would be more substantial and extensive media attention before that round, and thus, signals could be more informative than they would be if those states voted sequentially.

Naturally, this feature would improve the performance of any simultaneous or partly simultaneous system, relative to the sequential baseline case. The decisive question is whether plausible improvements in signal quality would be sufficient to overturn the dominance of the sequential system. To analyze this question, we compare the sequential benchmark with an additional set of simulations in which we increasing the precision of voter signals (i.e., $\sigma_\eta \downarrow$) in the simultaneous system and in the NASS system.

Even if we increase the precision of information in the simultaneous system by an unrealistically large extent, the sequential benchmark still does better: Even if we reduce the variance of the signal by 90%, welfare in the simultaneous system remains well below welfare in the sequential benchmark. Intuitively, this is the case because better information does not resolve the problem of vote-splitting that is endemic in the simultaneous system. Even if those voters who prefer the position that two of the candidates share knew candidate quality perfectly, they would still split their votes because they also have idiosyncratic preferences. Thus, even with nearly perfect information, there is a substantial probability that the Condorcet winner loses in a simultaneous primary.

As for the NASS system, we find that an increase in the precision of information can outperform the sequential benchmark. If σ_η is reduced by about 40 percent, the NASS system performs as well as the sequential benchmark in terms of welfare and would be superior in terms of the probability that the Condorcet winner wins the primary (for larger reductions in σ_η, the NASS proposal would be even more superior). The required reduction of σ_η by 40% is quite substantial, but not necessarily implausibly large.

A limitation of our analysis is that we do not know whether alternative coordination mechanisms would arise in a simultaneous primary to avoid vote splitting. There is clearly a large incentive to do that, for example using public opinion polls.[38] If this is the case, our simulations would underestimate the performance of a simultaneous primary system relative to a sequential one.

While this argument is theoretically valid, its practical impact may be limited. First, coordination in simultaneous primaries appears non-trivial to achieve in practice. In simultaneous party primaries for open seat state offices or U.S. Congress, there are often contests with several serious candidates who all receive substantial vote shares, and where the winner's vote share is substantially below 50 percent, indicating the potential importance of vote splitting.[39] This suggests that coordination facilitated by either candidates dropping out before the election or based on opinion polls cannot be taken for granted.

Second, coordination using informal mechanisms such as straw polls or opinion polls is likely to be based on substantially worse information quality than the outcome of an actual primary election in a state because the sample of people who participate in the straw poll or opinion poll is unlikely to be perfectly representative of the electorate. Also, attempts by the candidates to influence the coordination criterion in a way that is not reflective of true valence are more likely to be successful in straw polls than in statewide elections.[40] So, while informal coordination in simultaneous primaries might have the effect that the outcome in this system is not quite as bad as our simulations suggest, it is unlikely to change the qualitative result that simultaneous primaries perform worse than sequential primaries.

[38] Fey (1997) provides a formal model of how pre-electoral polls help coordination on the leading candidates by reducing vote-splitting in the election.

[39] For example, in the 2010 Republican primary for Governor of Illinois, five of the seven candidates received more than 14 percent of the votes each, and Bill Brady won with a vote share of just 20.3 percent. Moreover, only Brady came from "downstate", while the remaining (serious) candidates all came from the Chicago area, and there appears to have been considerable region-based vote-splitting – Brady received only 7 percent in Chicago and its suburbs, but won nevertheless because of his strong showing downstate and since the Chicago-based candidates split the vote there very evenly.

[40] Consider the Iowa Straw Poll, which is organized by the Republican party in the summer of the year before presidential nomination contests. A poor showing in the Iowa Straw Poll is often very problematic for a candidate and may effectively end his campaign (for example, in 2008, Tommy Thompson and Sam Brownback were effectively eliminated by this straw poll). For this reason, candidates often spend substantial resources in order to provide transportation or buy tickets for their supporters, diminishing the informational content of the voting outcome.

5 Conclusion

A fundamental question in the scientific analysis of politics is the performance of different political institutions. Our paper analyzes this question in a specific context, focusing on the trade-off between voter learning and voter coordination in different election systems that differ in the timing of voting. The U.S. Presidential Primary system provides a unique case study for these effects that are otherwise difficult to assess because we usually only observe voters make a decision once, at the end of the campaign when they vote.

At the start of presidential primaries, there are often several serious contenders. Some of them are likely to be ideologically closer substitutes for voters than others. In a simultaneous election with a large set of candidates, the candidate who would come out on top is not necessarily the Condorcet winner. In contrast, sequential elections allow voters to narrow down the field of contenders as a way of avoiding vote-splitting among ideologically similar candidates. Sequential primaries therefore likely have facilitated the victory of candidates who were not the frontrunner at the beginning of the primary season, such as Obama (and possibly McCain) in 2008, and the very strong showing of Gary Hart in 1984.

We have presented a model of voting in sequential primaries based on the trade-off between coordination and learning about candidate quality. From a theoretical perspective, the coordination afforded by sequential elections may be beneficial or detrimental. While sequential elections allow voters to coordinate and thus avoid that a candidate wins just because his ideological opponents split the votes of their supporters among each other, their disadvantage is that, once coordination has occurred, there is little possibility to correct an error made in early elections, as 'momentum' dominates. Moreover, our empirical results show that the probability of the full information Condorcet winner dropping out after the first few primaries is substantial.

Nevertheless, sequential elections dominate simultaneous ones if valence differences between candidates are not too large relative to the importance of position differences; if the signal quality in early states is high; and if there is substantial vote-splitting between ideologically similar candidates. In contrast, when valence differences are important relative to the importance of position differences, vote-splitting is not too important and the signal quality is bad, then a simultaneous primary system may be superior.

We estimate the model using data from the 2000–2008 Democratic primaries, and the 2000, 2008 and 2012 Republican ones, and use the parameter estimates to evaluate the relative performance of different temporal organizations of the primaries. Our results suggest that vote-splitting would be a severe problem in a simultaneous primary system, and that it therefore would perform substantially worse from a welfare point of view than the current sequential system.

In contrast, a reform proposal by the National Association of Secretaries of State comes relatively close to the sequential benchmark system with respect to welfare in our simulations. The NASS proposal also has the advantage – not modeled in our analysis – that it might be perceived as fairer, as the sequence in which states in the East, Midwest,

South and West regions would rotate, and so the identity of the more important early states would change.[41] If the NASS system leads to more informative signals, because of the increased stakes of each round, it would dominate the status quo arrangement. Moreover, a robust finding of our analysis is that, if the national parties were to use early election results to winnow the field of contenders earlier, rather than wait for completely voluntary exit, then the performance of the Presidential primary system would improve.

Finally, while U.S. presidential primaries are carried out sequentially, state-level contests (e.g. gubernatorial and U.S. Senate primaries) are universally organized as simultaneous elections. Of course, one could imagine to organize gubernatorial primaries as sequential county-by-county elections, for example. A positive consequence would be that a sequential organization would help to avoid vote splitting in cases such as the one described in footnote 39 above. On the other hand, modern campaigning with its emphasis on media advertising probably requires that all elections in one media market take place simultaneously, and the number of significant media markets is very small in most states, so there might be practical limits to implementing a sequential election system at the state level.

[41] Moreover, the NASS proposal is relatively easy to implement, while a completely sequential primary system with one state voting per week would take about a year to complete.

6 Appendix

6.1 Proof of Proposition 1

1. $\mu < 1/2$**.** Since $1-\mu > 1/2$, Candidate 1 receives an absolute majority of votes in every district, whether he competes against one or two opponents. The election system only affects whether the votes of type $\theta = 1$ voters are split or united.

2(a). $\mu \in (1/2, 2/3)$ **and sequential elections.** Candidate 2 gets more votes in the first district than Candidate 3 if and only if $v_2 + \eta_2^1 > v_3 + \eta_3^1$. Since $\eta_3 - \eta_2$ is distributed according to $N(0, 2\sigma_\eta^2)$, for given v_2 and v_3, Candidate 2 wins with probability $\Phi\left(\frac{v_2 - v_3}{\sqrt{2}\sigma_\eta}\right)$. Note that $v_2 - v_3$ is distributed according to $N(0, 2\sigma_v^2)$. Without loss of generality, we can focus on the case $v_2 > v_3$; conditioning on this event, the density of $v_2 - v_3$ is given by $2\phi\left(\frac{t}{\sqrt{2}\sigma_v}\right)$. Thus, the probability that the better candidate wins is given by

$$2\int_0^\infty \Phi\left(\frac{t}{\sqrt{2}\sigma_\eta}\right)\phi\left(\frac{t}{\sqrt{2}\sigma_v}\right)dt = \sqrt{2}\sigma_v\left[1 - \frac{\arctan\left(\frac{\sigma_\eta}{\sigma_v}\right)}{\pi}\right]. \tag{14}$$

Since the arc tan is an increasing function and lies between 0 and π (for positive arguments, such as here), it is easy to see that this probability is decreasing in σ_η and increasing in σ_v.

2(b). $\mu \in (1/2, 2/3)$ **and simultaneous elections.** It is useful to denote by ϕ_α, $\alpha \in \{v, \eta, \varepsilon\}$, the probability density function of the normal distribution of variable α. The voters in district s observe signal $Z_j^s = v_j + \eta_j^s$. Using Bayes' rule, the updated expected value of Candidate j's valence is

$$\hat{v}_j^s = \int_{-\infty}^{\infty} \frac{\phi_v(t)\phi_\eta(Z_j^s - t)}{\int_{-\infty}^{\infty}\phi_v(t')\phi_\eta(Z_j^s - t')dt'} t dt = \frac{{\sigma_v}^2}{{\sigma_v}^2 + {\sigma_\eta}^2} Z_j^s. \tag{15}$$

If Voter i in district s has type $\theta = 1$, he votes for Candidate 2 if $\hat{v}_2^s + \varepsilon_2^i > \hat{v}_3^s + \varepsilon_3^i$, and for Candidate 3 otherwise. Rearranging, the percentage of type $\theta = 1$ voters who vote for Candidate 2 is equal to

$$\text{Prob}(\varepsilon_3 - \varepsilon_2 \leq \hat{v}_2^s - \hat{v}_3^s) = \text{Prob}\left(\varepsilon_3 - \varepsilon_2 \leq \frac{{\sigma_v}^2}{{\sigma_v}^2 + {\sigma_\eta}^2}[Z_2^s - Z_3^s]\right) = \Phi\left(\frac{{\sigma_v}^2[Z_2^s - Z_3^s]}{\sqrt{2}\sigma_\varepsilon({\sigma_v}^2 + {\sigma_\eta}^2)}\right). \tag{16}$$

Similarly, Candidate 3's share of the vote of $\theta = 1$ types is equal to

$$1 - \Phi\left(\frac{{\sigma_v}^2[Z_2^s - Z_3^s]}{\sqrt{2}\sigma_\varepsilon({\sigma_v}^2 + {\sigma_\eta}^2)}\right).$$

Candidate 1, the Condorcet loser, receives all votes from $\theta = 0$ types (a proportion $1-\mu$ of the electorate) and wins a particular district s if and only if

$$1 - \mu > \mu \cdot \max\left(\Phi\left(\frac{{\sigma_v}^2[Z_2^s - Z_3^s]}{\sqrt{2}\sigma_\varepsilon({\sigma_v}^2 + {\sigma_\eta}^2)}\right), 1 - \Phi\left(\frac{{\sigma_v}^2[Z_2^s - Z_3^s]}{\sqrt{2}\sigma_\varepsilon({\sigma_v}^2 + {\sigma_\eta}^2)}\right)\right), \tag{17}$$

hence if

$$\frac{2\mu - 1}{\mu} < \Phi\left(\frac{{\sigma_v}^2[Z_2^s - Z_3^s]}{\sqrt{2}\sigma_\varepsilon({\sigma_v}^2 + {\sigma_\eta}^2)}\right) < \frac{1-\mu}{\mu}. \tag{18}$$

Denoting the inverse of the cumulative distribution of the standard normal distribution by Φ^{-1}, and letting $\kappa = \frac{\sigma_v^2}{\sqrt{2}\sigma_\varepsilon(\sigma_v{}^2+\sigma_\eta{}^2)}$, we can write this as

$$\Phi^{-1}\left(\frac{2\mu-1}{\mu}\right) < \kappa(v_2 - v_3) + \kappa(\eta_2 + \eta_3) < \Phi^{-1}\left(\frac{1-\mu}{\mu}\right) \tag{19}$$

For given v_2 and v_3, the term in the middle is normally distributed with expected value $\kappa(v_2 - v_3)$ and variance $2\kappa^2\sigma_\eta{}^2$. Thus, the percentage of districts won by Candidate 1 is given by

$$\begin{aligned}\text{Prob}\left(\Phi^{-1}\left(\frac{2\mu-1}{\mu}\right) - \kappa(v_2 - v_3) < \kappa(\eta_2 - \eta_3) < \Phi^{-1}\left(\frac{1-\mu}{\mu}\right) - \kappa(v_2 - v_3)\right) = \\ \Phi\left(\frac{\Phi^{-1}\left(\frac{1-\mu}{\mu}\right) - \kappa(v_2 - v_3)}{\sqrt{2}\kappa\sigma_\eta}\right) - \Phi\left(\frac{\Phi^{-1}\left(\frac{2\mu-1}{\mu}\right) - \kappa(v_2 - v_3)}{\sqrt{2}\kappa\sigma_\eta}\right) = \\ \Phi\left(\frac{\Phi^{-1}\left(\frac{1-\mu}{\mu}\right) - \kappa(v_2 - v_3)}{\sqrt{2}\kappa\sigma_\eta}\right) - \Phi\left(\frac{-\Phi^{-1}\left(\frac{1-\mu}{\mu}\right) - \kappa(v_2 - v_3)}{\sqrt{2}\kappa\sigma_\eta}\right),\end{aligned} \tag{20}$$

where the last inequality uses the fact that $\Phi^{-1}\left(\frac{2\mu-1}{\mu}\right) = -\Phi^{-1}\left(\frac{1-\mu}{\mu}\right)$, because $\frac{2\mu-1}{\mu}$ and $\frac{1-\mu}{\mu}$ are symmetric around $1/2$ (i.e., add up to 1).

Again, suppose that $v_2 > v_3$, so that Candidate 2 is the toughest competitor for the nomination. The percentage of districts won by Candidate 2 is

$$\Phi\left(\frac{-\Phi^{-1}\left(\frac{1-\mu}{\mu}\right) - \kappa(v_2 - v_3)}{\sqrt{2}\kappa\sigma_\eta}\right). \tag{21}$$

Candidate 1 wins the nomination if (20) is larger than (21) he wins more districts than Candidate 2, hence if

$$\Phi\left(\frac{\Phi^{-1}\left(\frac{1-\mu}{\mu}\right) - \kappa(v_2 - v_3)}{\sqrt{2}\kappa\sigma_\eta}\right) > 2\Phi\left(\frac{-\Phi^{-1}\left(\frac{1-\mu}{\mu}\right) - \kappa(v_2 - v_3)}{\sqrt{2}\kappa\sigma_\eta}\right). \tag{22}$$

Note that the left hand side is decreasing in μ, while the right hand side is increasing in μ. Thus, if (22) holds for a particular level of μ, then it also holds for all smaller levels of μ (equivalently, all higher levels of $1-\mu$). This is intuitive, since $1-\mu$ is the percentage of voters who support Candidate 1. Let μ^* denote the level of μ such that (22) holds with equality.

Consider first the case of $\mu = 1/2$, such that $\frac{1-\mu}{\mu} = 1$ and hence $\Phi^{-1}\left(\frac{1-\mu}{\mu}\right) = \infty$. Clearly, (22) holds, as the left hand side goes to 1, while the right hand side goes to 0. Intuitively, if $\mu = 1/2$, then any sort of vote-splitting between Candidates 2 and 3 guarantees that Candidate 1 wins all districts. Since both sides are continuous in μ, the same result holds (for any given v_2 and v_3) for μ sufficiently close to $1/2$. Now consider the case of $\mu = 2/3$, such that $\frac{1-\mu}{\mu} = 1/2$. Since $\Phi^{-1}(1/2) = 0$, (22) is clearly violated.

Consider now the effect of changes in σ_ε, σ_η and σ_v on (22). Note first that $\kappa = \frac{\sigma_v^2}{\sqrt{2}\sigma_\varepsilon(\sigma_v{}^2+\sigma_\eta{}^2)}$ is decreasing in σ_ε and increasing in σ_v. Furthermore, the left hand side of (22) is decreasing in κ (as $(1-\mu)/\mu > 1/2$, and thus $\Phi^{-1}(\frac{1-\mu}{\mu}) > 0$), while the right hand side is increasing in κ by the same argument. Thus, to preserve equality between the two sides of (22), an increase of κ needs to be balanced by a decrease of μ^*. Consequently, μ^* decreases in σ_v, and increases in σ_ε.

We now analyze the effect of σ_η. Consider the difference of the left-hand and right-hand side of (22), and substitute for κ and set the expression equal to 0 (which implicitly determines the value of μ^*); this yields

$$Z = \Phi\left(\frac{\Phi^{-1}\left(\frac{1-\mu}{\mu}\right)}{\frac{\sigma_v^2\sigma_\eta}{\sigma_\varepsilon(\sigma_v^2+\sigma_\eta^2)}} - \frac{v_2-v_3}{\sqrt{2}\sigma_\eta}\right) - 2\Phi\left(\frac{-\Phi^{-1}\left(\frac{1-\mu}{\mu}\right)}{\frac{\sigma_v^2\sigma_\eta}{\sigma_\varepsilon(\sigma_v^2+\sigma_\eta^2)}} - \frac{v_2-v_3}{\sqrt{2}\sigma_\eta}\right) = 0. \tag{23}$$

Since $\Phi(\cdot)$ is an increasing function, $\Phi^{-1}\left(\frac{1-\mu}{\mu}\right)$ is decreasing in μ, and thus $\frac{\partial Z}{\partial \mu}$. Consequently, the sign of

$$\frac{d\mu^*}{\sigma_\eta} = -\frac{\frac{\partial Z}{\partial \sigma_\eta}}{\frac{\partial Z}{\partial \mu}}$$

is the same as the sign of $\frac{\partial Z}{\partial \sigma_\eta}$. We have

$$\begin{aligned}\frac{\partial Z}{\partial \sigma_\eta} = &\left[\phi\left(\frac{\Phi^{-1}\left(\frac{1-\mu}{\mu}\right)}{\frac{\sigma_v^2\sigma_\eta}{\sigma_\varepsilon(\sigma_v^2+\sigma_\eta^2)}} - \frac{v_2-v_3}{\sqrt{2}\sigma_\eta}\right) + 2\phi\left(\frac{\Phi^{-1}\left(-\frac{1-\mu}{\mu}\right)}{\frac{\sigma_v^2\sigma_\eta}{\sigma_\varepsilon(\sigma_v^2+\sigma_\eta^2)}} - \frac{v_2-v_3}{\sqrt{2}\sigma_\eta}\right)\right] \times \left[\frac{\sigma_\varepsilon}{\sigma_v^2}\Phi^{-1}\left(\frac{1-\mu}{\mu}\right)\frac{\sigma_\eta^2-\sigma_v^2}{\sigma_\eta^2}\right] + \\ &\left[\phi\left(\frac{\Phi^{-1}\left(\frac{1-\mu}{\mu}\right)}{\frac{\sigma_v^2\sigma_\eta}{\sigma_\varepsilon(\sigma_v^2+\sigma_\eta^2)}} - \frac{v_2-v_3}{\sqrt{2}\sigma_\eta}\right) - 2\phi\left(\frac{\Phi^{-1}\left(-\frac{1-\mu}{\mu}\right)}{\frac{\sigma_v^2\sigma_\eta}{\sigma_\varepsilon(\sigma_v^2+\sigma_\eta^2)}} - \frac{v_2-v_3}{\sqrt{2}\sigma_\eta}\right)\right] \times \frac{v_2-v_3}{\sqrt{2}\sigma_\eta^2}\end{aligned} \tag{24}$$

(24) is greater than

$$2\phi\left(\frac{\Phi^{-1}\left(-\frac{1-\mu}{\mu}\right)}{\frac{\sigma_v^2\sigma_\eta}{\sigma_\varepsilon(\sigma_v^2+\sigma_\eta^2)}} - \frac{v_2-v_3}{\sqrt{2}\sigma_\eta}\right)\left[\frac{\sigma_\varepsilon}{\sigma_v^2}\Phi^{-1}\left(\frac{1-\mu}{\mu}\right)\frac{\sigma_\eta^2-\sigma_v^2}{\sigma_\eta^2} - \frac{v_2-v_3}{\sqrt{2}\sigma_\eta^2}\right]$$

Since the term in square brackets goes to $\frac{\sigma_\varepsilon}{\sigma_v^2}\Phi^{-1}\left(\frac{1-\mu}{\mu}\right) > 0$ for $\sigma_\eta \to \infty$, (24) is positive for σ_η sufficiently large. Thus, for σ_η sufficiently large, $\frac{d\mu^*}{d\sigma_\eta}$ is positive. In contrast, for $v_2 = v_3$ and $\sigma_\eta < \sigma_v$, (24) and hence $\frac{d\mu^*}{d\sigma_\eta}$ is negative.

3. $\mu > 2/3$**.** In this case, Candidate 1 receives less than a third of the votes in every district, so that he loses in every district. Without loss of generality, suppose again that $v_2 > v_3$.

Under simultaneous elections, Candidate 2 wins in district s if

$$v_2 + \eta_2^s > v_3 + \eta_3^s. \tag{25}$$

Thus, for a given $v_2 > v_3$, the proportion of districts won by Candidate 2 is equal to $\Phi\left(\frac{v_2-v_3}{\sqrt{2}\sigma_\eta}\right) > 1/2$. Consequently, Candidate 2 is certain to win the nomination contest.

Under sequential elections, the winner of the first district (either Candidate 2 or Candidate 3) gets a vote share μ in all following districts and thus wins the nomination. The probability that Candidate 2 is the winner of the first district is the same as in (14) in Case 2 above. Thus, the better candidate is likely to win the nomination, but there is a positive probability that the other candidate with the same policy position wins instead. ■

6.2 Updating about candidate quality

Proposition 2 *The vote shares of candidates in state s satisfy the following equation system:*

$$\begin{aligned}
W_j^s &= (1-\mu^s)\int_{-\infty}^{\infty}\prod_{J_0^s\setminus\{j\}}\Phi\left(\frac{\hat{v}_j^s-\hat{v}_{j'}^s+\varepsilon_j}{\sigma_\varepsilon}\right)\cdot\prod_{J_1^s}\Phi\left(\frac{\lambda+\varepsilon_j+\hat{v}_j^s-\hat{v}_{j'}^s}{\sigma_\varepsilon}\right)\cdot\phi_\varepsilon(\varepsilon_j)d\varepsilon_j + \\
&\quad \mu^s\int_{-\infty}^{\infty}\prod_{J_0^s\setminus\{j\}}\Phi\left(\frac{\hat{v}_j^s-\hat{v}_{j'}^s+\varepsilon_j}{\sigma_\varepsilon}\right)\cdot\prod_{J_1^s}\Phi\left(\frac{-\lambda+\varepsilon_j+\hat{v}_j^s-\hat{v}_{j'}^s}{\sigma_\varepsilon}\right)\cdot\phi_\varepsilon(\varepsilon_j)d\varepsilon_j\,,\forall\, j\in J_0^s \\
W_j^s &= (1-\mu^s)\int_{-\infty}^{\infty}\prod_{J_0^s}\Phi\left(\frac{-\lambda+\hat{v}_j^s-\hat{v}_{j'}^s+\varepsilon_j}{\sigma_\varepsilon}\right)\cdot\prod_{J_1^s\setminus\{j\}}\Phi\left(\frac{\varepsilon_j+\hat{v}_j^s-\hat{v}_{j'}^s}{\sigma_\varepsilon}\right)\cdot\phi_\varepsilon(\varepsilon_j)d\varepsilon_j + \\
&\quad \mu^s\int_{-\infty}^{\infty}\prod_{J_0^s}\Phi\left(\frac{\lambda+\hat{v}_j^s-\hat{v}_{j'}^s+\varepsilon_j}{\sigma_\varepsilon}\right)\cdot\prod_{J_1^s\setminus\{j\}}\Phi\left(\frac{\varepsilon_j+\hat{v}_j^s-\hat{v}_{j'}^s}{\sigma_\varepsilon}\right)\cdot\phi_\varepsilon(\varepsilon_j)d\varepsilon_j\,,\forall\, j\in J_1^s
\end{aligned} \tag{26}$$

There exists a unique vector of valence values $(0, x_2, x_3, \ldots x_k)$ *such that all solutions of* (26) *are of the form* $(0, x_2, x_3, \ldots, x_k)+$ $(c, c, \ldots, c)$, $c \in \mathbb{R}$.

Proof. For a given ε_j, (3) is satisfied if and only if

$$\varepsilon_{j'} < \hat{v}_j^s - \hat{v}_{j'}^s + \varepsilon_j - \lambda[d(j,\theta) - d(j',\theta)] \text{ for all } j' \neq j\,. \tag{27}$$

First consider a voter of type $\theta = 0$. Since the ε's are distributed independently, the probability that such a voter votes for Candidate j is

$$\prod_{J_0^s\setminus\{j\}}\Phi\left(\frac{\hat{v}_j^s-\hat{v}_{j'}^s+\varepsilon_j}{\sigma_\varepsilon}\right)\cdot\prod_{J_1^s}\Phi\left(\frac{\lambda+\varepsilon_j+\hat{v}_j^s-\hat{v}_{j'}^s}{\sigma_\varepsilon}\right). \tag{28}$$

Integrating over the possible realizations of ε_j shows that the proportion of type 0 voters who vote for Candidate $j \in J_0^s$ is

$$\int_{-\infty}^{\infty}\prod_{J_0^s\setminus\{j\}}\Phi\left(\frac{\hat{v}_j^s-\hat{v}_{j'}^s+\varepsilon_j}{\sigma_\varepsilon}\right)\cdot\prod_{J_1}\Phi\left(\frac{\lambda+\varepsilon_j+\hat{v}_j^s-\hat{v}_{j'}^s}{\sigma_\varepsilon}\right)\cdot\phi_\varepsilon(\varepsilon_j)d\varepsilon_j. \tag{29}$$

Similarly, the share of type 1 voters who vote for Candidate j is

$$\int_{-\infty}^{\infty}\prod_{J_0^s\setminus\{j\}}\Phi\left(\frac{\hat{v}_j^s-\hat{v}_{j'}^s+\varepsilon_j}{\sigma_\varepsilon}\right)\cdot\prod_{J_1^s}\Phi\left(\frac{-\lambda+\varepsilon_j+\hat{v}_j^s-\hat{v}_{j'}^s}{\sigma_\varepsilon}\right)\cdot\phi_\varepsilon(\varepsilon_j)d\varepsilon_j. \tag{30}$$

The total vote share of Candidate $j \in J_0^s$ is then given by the weighted average of (29) and (30), where the weights are $(1-\mu^s)$ and μ^s. In an analogous way, the total vote share of Candidate $j \in J_1^s$ can be derived. Thus, the vote shares of candidates are given by (26).

The claim of Proposition 2 is that there exists a unique vector of valence values $(0, x_2, x_3, \ldots x_k)$ such that all solutions of (26) are of the form $(0, x_2, x_3, \ldots, x_k) + (c, c, \ldots, c)$, $c \in \mathbb{R}$.

Existence follows by construction: Since the vector W^r is generated using the realized vector of estimated valences $(\hat{v}_j^r)_{j=1,\ldots,k}$, a solution to (26) exists.

Furthermore, it is clear that any vector of the form $(0, x_2, x_3, \ldots, x_k) + (c, c, \ldots, c)$ also satisfies (26). It remains to be shown that there cannot be a solution of the form $(0, y_2, y_3, \ldots, y_k)$ with $(0, y_2, y_3, \ldots, y_k) \neq (0, x_2, x_3, \ldots, x_k)$. Assume to the contrary, and let $\bar{k}$ be the candidate for whom $y_j - x_j$ is maximal. If $y_{\bar{k}} - x_{\bar{k}} > 0$, then substituting in the corresponding equation of (26) shows that candidate $\bar{k}$ receives a strictly higher vote share than $W^r_{\bar{k}}$, a contradiction. Similarly, let $\underline{k}$ be the candidate for whom $y_j - x_j$ is minimal. If $y_{\underline{k}} - x_{\underline{k}} < 0$, then substituting in the corresponding equation of (26) shows that candidate $\underline{k}$ receives a strictly smaller vote share than $W^r_{\underline{k}}$, a contradiction. But then, it must be true that $y_j = x_j$ for all $j = 2, \ldots, k$. This completes the proof of Proposition 2. ■

Note that vote shares are determined only by the *difference* between the candidates' estimated valences, so we can only determine those differences. However, it is also immaterial which of these possible beliefs a voter in a later state uses to infer the signals observed by the voters of that state.

Corollary 1 *Given a set of ex-post valence beliefs* $(0, x_2, x_3, \ldots, x_k) + (c, c, \ldots, c)$*,* $c \in \mathbb{R}$*, there is a unique vector of signals* $(0, y_2, y_3, \ldots, y_k)$ *such that all solutions to the system of equations given in* (4)*, for* $j \in \{1, \ldots, k\}$*, are of the form* $(0, y_2, y_3, \ldots, y_k) + (\gamma, \gamma, \ldots, \gamma)$*.*

Proof. This follows from the fact that equations (4) form a linear system in ex-post valences and observed signals for all candidates competing in state s. ■

By observing vote shares in the election of a prior state, a voter can infer signals up to a constant. As already pointed out, voters determine their preferred candidate on the basis of differences in ex-post perceived valence, and these differences are determined by differences in the valence signals observed by voters of the state. In other words, a uniform shift of the ex-ante beliefs about all candidates by c translates into a uniform shift of the ex-post beliefs (i.e., after the state-specific signal), leaving the difference between the valence estimates for the different candidates, and hence the voter's voting decision, unaffected. The value of γ is immaterial in determining voting shares and can be normalized to zero.

Finally, note that the right-hand sides of (26) are homogeneous of degree 0 in $(\varepsilon, \hat{v}^s, \sigma_\varepsilon)$. It is therefore useful to normalize $\sigma_\varepsilon \equiv 1$. Thus, all other parameters in the model are effectively expressed as multiples of the standard deviation of the idiosyncratic preference shock ε.

6.3 Estimation Algorithm Details

The estimation algorithm proceeds as follows for the Democratic primaries. Consider a given set of parameter values, $\tilde{\sigma}_v$, $\tilde{\sigma}_\eta$, $\tilde{\lambda}$, and $\tilde{S}_\mu$. These parameter values will be the initial values at the start of the algorithm, or intermediate values given by the Newton-Raphson optimization routine while the algorithm is in progress. We draw a set of R vectors each contain 11 normally distributed valence draws, v^r_j, with $r = 1, \ldots R$ and $j = 1, \ldots 11$ ($R = 50,000$, resulting in small sampling errors). These draws have mean zero and standard deviation $\tilde{\sigma}_v$. For each v^r_j we draw a sequence of

normally distributed noise terms, $\eta^r_{j,s}$ which we use to form signals $Z^r_{j,s} = v^r_j + \eta^r_{j,s}$. For draws $j = 1$ and 2, which are assigned to the candidates in the 2000 primary, the noise sequences (and associated signals) consist of 15 noise draws each, because the 2000 primary took place in 15 states. For valence draws indexed draws $j = 3, \ldots 8$, which are assigned to the candidates in the 2004 primary, the noise sequences have length of 28, while for valence draws $j = 9 \ldots 11$, the noise (and signal) sequences have length of 50.[42] The noise and signal sequences are assigned to states as follows. For each primary, we order states by the round of voting (early rounds first), and then alphabetically within round (any sort within round would be fine, as we do not use state characteristics). The first state in year y takes the first noise/signal from the corresponding sequence, the second state takes the second noise/signal, and so on until the sequence is exhausted.

The valence draws and the signal sequences are used to obtain perceived valence sequences, $\hat{v}^r_{j,s}$ using the Bayesian updating expressions reported in the theory section of this paper. These are assigned to candidates as follows. Bradley is assigned at random either the sequence indexed $j = 1$ or the sequence indexed $j = 2$. Gore is assigned the other sequence. Sequences indexed $j = 3, 4, 5$ are assigned to the three "outsider" candidates of the 2004 primary as follows. For replication r, the sequence $\hat{v}^r_{j,s}$ is assigned to Clark if it is the lowest of the three perceived valences after 14 signal draws. Of the remaining two sequences, the one corresponding to the lowest perceived valence after 16 signal draws is assigned to Dean. The remaining sequence is assigned to Edwards. Sequences indexed $j = 6, 7, 8$ are assigned to the insider candidates of the 2004 primary as follows. For replication r, among the sequences $j = 6, 7, 8$, the one with the lowest of perceived valence after one signal after one signal is discarded (i.e., assigned Gephardt). From the surviving sequences, the one with the highest perceived valence after 9 signals is assigned to Kerry, while the other is assigned to Lieberman. Sequences indexed $j = 9, 10$ are assigned to Edwards/Obama as follows. For replication r, the sequence that corresponds to the lowest perceived valence after five signals is assigned to Edwards, while the other sequence is assigned to Obama. Finally, the last sequence, indexed by $j = 11$, is assigned to Hilary Clinton.

We also generate R sequences of μ^r_s each with 15 elements, which are assigned to the 15 states taking place in the 2000 Democratic primary. Each element of the sequence is an i.i.d. draw from a uniform distribution with mean 0.5 and support $\tilde{S}_\mu$. Similarly, we generate R sequences of μ^r_s each with 27 elements, which are assigned to the 27 states in the 2004 Democratic primary that are used in the estimation, and another R sequences of μ^r_s each with 50 elements which are assigned to the 50 states in the 2008 Democratic primary.[43] Following this, we compute the vote share of the insider candidate predicted by the theoretical model, $W((v^s_y)^r, (\eta_s)^r, (\vec{\eta}_{\rho<\rho_s})^r, (\mu^s_y)^r, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{\sigma}_v)$, for each replication r and for each of the $N_{obs} = 92$ election contests in our data.[44] The average over all R replications yields the simulation-based counter-parts of $E_{v,\vec{\mu},\vec{\eta}}\left[W(v^s_y, \eta_s, \vec{\eta}_{\rho<\rho_s}, \mu^s_y, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{\sigma}_v)|y, s\right]$ for each year-state pair. These predicted vote share are

[42] There are 29 states in the 2004 primary, but Vermont in the last round is dropped. There are 50 states in the 2008 primary.

[43] There is no reason to draw a μ^r_s for Iowa in the 2004 primary, because that state was used only for updating valences and not for its vote shares. Iowa shares were not used because two major candidates decided to skip it in 2004. We assigned the sequences of μ^r_s to states in the same manner that we assigned the signal sequences. Given that μ draws are iid and that outcomes in state s do not depend on the μs of prior states, a random allocation would have been fine.

[44] The integral in the expression for the vote share is obtained numerically as follows. The distribution of ϵ is discretized and evaluated at 50 equally spaced points between -3 and 3, and the the sum of the probabilities adjusted to sum to unity.

"matched" with the actual (observed) vote shares of the corresponding state-year pairs, W_y^s, with states ordered within a year in the same fashion as above (first by round, then alphabetically). Thus, the simulation-based sample-analog counterpart of moment condition (7) is:

$$\frac{1}{N_{obs}} \sum_{y,s} \left\{ \left(\frac{1}{R} \sum_{r} \left[W((v_y^s)^r, (\eta_s)^r, (\vec{\eta}_{\rho<\rho_s})^r, (\mu_y^s)^r, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{\sigma}_v) \right] - W_y^s \right) \mathbf{1}_{(s,y)\in g} \right\} = 0 \tag{31}$$

The simulation-based analog counterpart of moment condition (8) involves a small modification of the above, and given by

$$\frac{1}{N_{obs}} \sum_{y,s} \left\{ \left(\frac{1}{R} \sum_{r} [\max\{W(\cdot), 1 - W(\cdot)\}] - \max\{W_y^s, 1 - W_y^s\} \right) \mathbf{1}_{(s,y)\in g} \right\} = 0 \tag{32}$$

where the arguments of $W(\cdot)$ are as in the sample-analog moment condition (31).

We obtain the remaining simulation-based moment conditions in a similar manner. Using the predicted vote shares for replication r for each state, we compute the (partial) average vote share, $(\tilde{W}_g)^r$, of the insider candidate for each state contest group g and replication r. The absolute difference between $W((v_y^s)^r, (\eta_s)^r, (\vec{\eta}_{\rho<\rho_s})^r, (\mu_y^s)^r, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{\sigma}_v)$ and $(\tilde{W}_g)^r$ is averaged over all R replications to yield the simulation-based counterpart of $E_{v,\vec{\mu},\vec{\eta}}\left[\left| W(v_y^s, \eta_s, \vec{\eta}_{\rho<\rho_s}, \mu_y^s, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{\sigma}_v) - \tilde{W}_{g_{y,s}} \right| \mid y, s \right]$ for each year-state pair. These are then matched with the observed values of $|W_y^s - \bar{W}_{g_{y,s}}|$. Thus, the simulation-based sample-analog counterpart of moment condition (9) is:

$$\frac{1}{N_{obs}} \sum_{y,s} \left\{ \left(\frac{1}{R} \sum_{r} \left[\left| W((v_y^s)^r, (\eta_s)^r, (\vec{\eta}_{\rho<\rho_s})^r, (\mu_y^s)^r, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{\sigma}_v) - (\tilde{W}_{g_{y,s}})^r \right| \right] - \left| W_y^s - \bar{W}_{g_{y,s}} \right| \right) \mathbf{1}_{(s,y)\in g} \right\} = 0 \tag{33}$$

The last two simulation-based moments are obtained in an analogous fashion. This completes the description of how the simulation-based counter-parts of the 15 moments conditions are obtained. We next turn to the procedure through which we obtain the parameter estimates.

Let θ be the vector of all four unknown parameters to be estimated. Then, denote by $g(W_y^s, \theta)$ the column vector of the terms inside the outer summation (or expectation) of the moment conditions that correspond to state s in year y. This vector takes a different value for each of the $N_{obs} = 92$ observations in the dataset. Denote the 15 element column vector of the sample averages of $g(W_y^s, \theta)$ over all observations by $\bar{g}(\mathbf{W}, \theta)$, where $\mathbf{W}$ is the vector of vote shares. With the number of parameters being fewer than the number of equations formed by these sample analogs, it is not possible to find values of θ such that $\bar{g}(\mathbf{W}, \theta) = 0$ for all 15 elements. The (Simulated) Generalized Method of Moments approach attempts to find values of θ that make the right hand side of the preceding equation system "as close to zero" as possible. With $\bar{g}(\mathbf{W}, \theta)$ being a vector, the definition of "close" depends on the norm used and on the relative weight of each element of the vector. Given that units for all moments are similarly scaled, reflecting changes in vote percentages (recall that higher order moments are in standard deviations rather than variances), we adopt the following simple procedure for our base results. We find values of θ that minimize

$$Q(\theta) = \bar{g}(\mathbf{W}, \theta)' \bar{g}(\mathbf{W}, \theta) \tag{34}$$

that is, we find parameter values that minimize the unweighted Euclidean norm of the elements of $\bar{g}(\mathbf{W}, \theta)$ (the squared differences between the elements of $\bar{g}(\mathbf{W}, \theta)$ and zero). These parameter values, $\hat{\theta}$, are consistent estimates of θ.

Though moments are weighted equally in obtaining the parameter estimates, we weigh them differentially based to their informativeness when obtaining standard errors. In particular, let the 15-by-4 gradient matrix of the moment conditions evaluated at the parameter estimates be denoted by D, i.e., define $D = \frac{\partial \bar{g}(\mathbf{W},\hat{\theta})}{\partial \theta}$. We then obtain an estimate of the parameter co-variance matrix by using

$$\hat{\Omega} = \frac{1}{\sqrt{N_{obs}}} (D'D)^{-1} D' \hat{S} D (D'D)^{-1} \tag{35}$$

where $\hat{S}$ is a 15-by-15 matrix given by

$$\hat{S} = \frac{1}{N_{obs}} \sum_{y,s} g(W_y^s, \hat{\theta}) g(W_y^s, \hat{\theta})' \tag{36}$$

This procedure, which yields consistent estimates of the standard errors, is reminiscent of White's method in obtaining robust standard errors in OLS regressions. One could also obtain estimates in a procedure that is reminiscent of GLS, by weighing the moments in the estimation step using the $\hat{S}$ matrix. The objective function for this procedure is

$$Q_{CW}(\theta) = \bar{g}(\mathbf{W}, \theta)' S(\theta)^{-1} \bar{g}(\mathbf{W}, \theta) \tag{37}$$

This approach continuously updates the moments as the parameters θ change. Like GLS, it is more efficient under the correct specification of the model, but has some weakness, such as a non-convex objective function and poor behavior at the "edges" of the parameter space.[45] The parameter estimates for the interior solution using this procedure yield similar values of S_μ and λ, but smaller value for σ_v and (particularly) σ_η. We attach less faith for these values because they tend to greatly underweight key moments used to match the reduction in the variability of vote shares. In any event, the simulation outcomes obtained with these values are within the range of those reported in our sensitivity analysis.

The estimation for the Republican primaries proceeds in an analogous fashion. There is one noteworthy difference. In the 2008 primary, Romney withdrew before Huckabee, even though many would argue that the former was "ahead" of the later. For this reason, we assign Huckabee to be the candidate with lowest actual valence among the two (finalist) conservatives running in that year, rather than the lowest perceived valence at the time of withdrawal. This permits the person who withdraws to sometimes be ahead of the candidate who stays.

[45] Yet another common procedure is to use a fixed $\hat{S}$ in equation 37 and iterate between equations 36 and 37 until some convergence criterion is satisfied. This procedure is reminiscent of the iterative approach to GLS estimation.

References

Ali, S. and N. Kartik (2012). Herding with collective preferences. *Economic Theory 51*(3), 601–626.

Anderson, S. P. and K. J. Meagher (2011). Primaries and presidents: Choosing a champion. Working paper.

Bartels, L. M. (1985). Expectations and preferences in presidential nominating campaigns. *The American Political Science Review 79*(3), 804–815.

Bartels, L. M. (1987). Candidate choice and the dynamics of the presidential nominating process. *American Journal of Political Science 31*(1), 1–30.

Bouton, L. and M. Castanheira (2012). One person, many votes: Divided majority and information aggregation. *Econometrica 80*, 43–87.

Bouton, L., M. Castanheira, and A. Llorente-Saguer (2012). Divided majority and information aggregation: Theory and experiment. CEPR Discussion Paper No. DP9234.

Callander, S. (2007). Bandwagons and momentum in sequential voting. *Review of Economic Studies 74*, 653–684.

Coughlin, P. (1992). *Probabilistic Voting Theory*. Cambridge, Cambridge University Press.

Dal Bo, E., P. Dal Bo, and J. Snyder (2009). Political dynasties. *Review of Economic Studies 76*, 115–142.

Degan, A. and A. Merlo (2006). Do voters vote sincerely? University of Pennsylvania.

Dekel, E. and M. Piccione (2000). Sequential voting procedures in symmetric binary elections. *Journal of Political Economy 108*(1), 34–55.

Deltas, G. and M. Polborn (2015). Candidate competition and voter learning in sequential primary elections: Theory and evidence. Working paper, University of Illinois.

Fey, M. (1997). Stability and Coordination in Duverger's Law: A Formal Model of Pre-Election Polls and Strategic Voting. *American Political Science Review 91*, 135–147.

Kawai, K. and Y. Watanabe (2010). Turnout and learning in sequential election: The case of U.S. presidential primaries. Working paper.

Kenny, P. J. and T. W. Rice (1994). The psychology of political momentum. *Political Research Quarterly 47*(4), 923–938.

Klumpp, T. and M. K. Polborn (2006). Primaries and the New Hampshire effect. *Journal of Public Economics 90*, 1073–1114.

Knight, B. and P. Hummel (2015). Sequential or simultaneous elections? An empirical welfare analysis. *International Economic Review 56*, 851–887.

Knight, B. and N. Schiff (2010). Momentum and social learning in presidential primaries. *Journal of Political Economy 118*, 1110–1150.

Krasa, S. and M. K. Polborn (2010). The binary policy model. *Journal of Economic Theory 145*(2), 661–688.

Lindbeck, A. and J. Weibull (1987). Balanced-budget redistribution as the outcome of political competition. *Public Choice 52*, 273–297.

Morton, R. and K. Williams (1999). Information Asymmetries and Simultaneous versus Sequential Voting. *American Political Science Review 93*, 51–67.

Morton, R. and K. Williams (2001). *Learning by Voting: Sequential Choices in Presidential Primaries and Other Elections*. University of Michigan Press.

Moulitsas, M. (2008a). Edwards voters' second choice, part II. Available at Daily Kos, http://www.dailykos.com/storyonly/2008/8/12/1101/20327/19/566444.

Moulitsas, M. (2008b). Wolfson's ridiculous speculation. Available at Daily Kos, http://www.dailykos.com/storyonly/2008/8/11/163149/067/156/566331.

Persson, T. and G. Tabellini (2000). *Political Economy: Explaining Economic Policy*. MIT Press.

Schwabe, R. (2010). Super Tuesday: Campaign Finance and the Dynamics of Sequential Elections. mimeo, Northwestern.

Stimson, K. (2008). The case for regional presidential primaries in 2012 and beyond. Available at http://nass.org/index.php?option=com_docman&task=doc_download&gid=165.