A Theory of Learning and Coordination in the Presidential Primary System

George Deltas∗ Helios Herrera† Mattias K. Polborn‡

May 8, 2010

Abstract

To analyze the performance of the U.S. presidential primary system, we develop a model with horizontally and vertically differentiated candidates. Voters are uncertain about candidates’ valences, and use results in earlier elections to update their beliefs. The temporal organization of primaries affects both voter learning and vote splitting (ie, several candidates in the same policy position competing for the same voter pool). Sequential voting minimizes vote-splitting in late districts, but voters may coordinate on the wrong candidate. We structurally estimate the model using the 2008 Democratic presidential primaries. Using the parameter estimates, we conduct policy experiments such as replacing the current system with simultaneous primaries and other proposed systems.

JEL Classification Numbers: D72, D60.

Keywords: Voting, Presidential primary elections, simultaneous versus sequential elections.

∗Department of Economics, University of Illinois, 1407 W. Gregory, Urbana, IL 61801 USA, E-mail: deltas@uiuc.edu
†SIPA, Columbia University and CIE-ITAM, Mexico, D.F. Email: hhh2108@columbia.edu
‡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

Candidates for the U.S. presidential election are determined through a sequence of elections within each political party, the primaries.¹ The nomination process is one of the most controversial institutions of America’s contemporary political landscape, as its sequential structure is perceived as inherently “unfair” because it shifts too much power to voters in early primary states. For this reason, 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. For the 2008 cycle, both the Democratic and the Republican National Committee chose rules that prohibited all but a few states to hold their primaries before February 5th. Florida and Michigan violated these rules and were punished by the DNC and RNC by taking away half of their delegates at the convention.² Thus, it appears that states have a strong interest in voting early (enough to risk such a punishment). Moreover, if the national parties’ decisions reflect an interest in the efficiency of the whole nomination process, then this behavior must be socially inefficient.

In order to evaluate these arguments, consider the following – only half-fictional – example of a nomination contest with three serious contenders at the time of the first elections, whom we call C, E and O. These candidates differ in some characteristics that are relevant for voters. First, candidate C has experience in Washington and would know on day 1 where the light switches are 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. In addition, there is uncertainty about the “valence” (e.g. absence of scandals, proneness of gaffes etc.) 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 coordinate their votes accordingly; also, in expectation of such coordination, the trailing candidate often drops out early. For example, if O gets more votes than E in the early elections, then even voters with ranking $E > O > C$ may vote for O, because they have determined that E has no chance of winning, and among the remaining relevant candidates, they prefer O. In this case, O will win the nomination if a majority of the electorate prefers him to C.

Such voter migrations between candidates are potentially crucial for election outcomes. 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

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 Throughout the primary process, the Democratic National committee even threatened to take away all of Florida’s and Michigan’s delegates, but then reduced the size of the penalty to one-half.
poll were Clinton 47, Obama 38, Edwards 11, while the actual election results were Obama 49.3, Clinton 47.1, Edwards 1.7. These numbers suggest that many Edwards supporters migrated to Obama, after Edwards dropped out of the race. Similarly, in a 12/26-30, 2007 poll by Opinion Research Corp for CNN (cited by Moulitsas (2008b)), 36% of Iowa Democrats polled state 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.

The benefit of a sequential system in our example is that the change voters are not split in most districts, thus increasing the likelihood that a change candidate wins (if this is the preference of a majority of voters). There is, however, a cost 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 later districts precisely because of coordination. The objective of our model is to provide a formal framework for the analysis of the trade-off between coordination and voter learning.

Because a reform of the primary system is frequently discussed, our question is not merely of theoretical interest, but also potentially policy-relevant. There are at least three different primary organization systems that have attracted considerable support among both commentators and politicians. The main alternatives to the current status quo of a sequential system appear to be a nationwide primary to be held on the same day, and a proposal by the National Association of the Secretaries of State for regional primaries. According to the latter 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 or utilitarian perspective.

Our theoretical model, set up in Section 3, provides some guidance as to which factors affect this trade-off. In Section 4, we characterize the optimal voting system for particularly simple limit cases where one of the two antagonistic effects outweighs the other. Yet, for more realistic settings, the question of the optimal voting system is a quantitative one. In Section 6, we estimate the structural parameters of our theoretical model using data from the 2008 Democratic primary. The estimated parameter values show that both key features of the theory (slow voter learning about candidate valence, and non-uniform substitutability of candidates with different political positions) are quantitatively important. In the first primary contest, the variability of the voters’ estimate of candidate valence is only about a third of the true valence variability (the reason is that signal quality is weak, and updating is thus not very responsive to the received signal in the first district). Moreover, the categorization of candidates into different “positions” appears to be very important for voters’ choice, if perceived valence differences are small. For example, if all candidates have the same perceived valence, they obtain more
than nine-tenths of the voters with the same political position, and less than a tenth of the other voters.

The estimated parameters feed directly into our institutional simulations in Section 7. All of our 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 of state voting. The first scenario assumes that all 50 states vote simultaneously; the second assumes that states vote sequentially and all three candidates remain in the race until the end; the third assumes that states vote sequentially but the weakest (in terms of voter perceptions) of the two candidates who share the same political position drops out after the fifth state votes.

Our results show that a sequential election with all candidates remaining in the race results in the highest expected valence and the highest probability that the Condorcet winner is elected, while a completely simultaneous election does worst. The two other setups yield intermediate results (though they are closer to the sequential setup). A simultaneous election makes the nomination of the sole candidate very probable, independent of this candidate’s valence, as vote-splitting between the two candidates in the same position is usually substantial and cross-over voting (i.e., voters with a preference for one position voting for a candidate in the other position) is only moderate. In the sequential election with all candidates staying in the race, there is some vote splitting in all districts, but the extent of it is sufficiently muted to be considerably less detrimental to the winning chances of the better of the two candidates in the same position. In the third scenario in which one of the two candidates who share the same position (namely, the one who is perceived as weaker by voters after the fifth district) drops out after the fifth state, the vote-splitting problem is reduced even further, but this comes at a substantial cost, as there is a substantial likelihood that the wrong candidate is eliminated (i.e., the candidate whose true valence is higher than the one of his competitor). Consequently, expected welfare decreases in this regime, relative to a completely sequential regime without dropout.

We also find that the optimal dropout time from a social point of view is quite late (approximately after 30 districts), but that the overwhelming part of the expected utility increase can be achieved by moving to a dropout after about 15 states. This is the reason why the NASS proposal does very well from a welfare point of view in our simulations. Assuming that all candidates stay in the race until after the first large regional contest, there are sufficiently many early elections to be relatively confident that the strongest candidates survive, yet vote splitting is absent in three out of four large regional contests. Relative to a primary structure modeled after the 2008 temporal structure that the Condorcet winner wins increases from 59.9% to 73.4%.

2 Related Literature

Several political science studies analyze the relation between voters’ expectations of which candidates will do well and their preference for these candidates. The study closest to our focus on the role of early primaries as a coordination device is Bartels (1987), which analyzes the 1984 Democratic presidential
primary. Bartels (1987, pp.13) describes the coordination process of those Democratic voters unhappy with the presumed front-runner, as follows.

At the beginning of the 1984 primary season, the question facing prospective voters was whether or not to support the obvious front-runner, Walter Mondale. Those who were most predisposed to support Mondale (on the basis of issue preferences […] ) would do so without undue soul-searching. On the other hand, a fair number of Democrats who were lukewarm (or worse) about Mondale’s candidacy may at least have entertained the possibility of supporting a different candidate. Their problem was to decide which alternative, if any, to turn to.

Having framed the problem in this way, we may ask ourselves what a prospective voter with an eye out for an alternative to Mondale would have been likely to know about the other candidates in the race. At the beginning of the campaign, the best answer is probably ‘very little’. But Hart’s second-place finish in Iowa, followed by his dramatic upset victory in New Hampshire changed that. By the end of February, our prospective voter was quite likely to know at least one thing about at least one challenger: that Gary Hart was out there, an alternative to Mondale with significant popular support, [suggesting that] a vote for Hart would not be wasted.

In the empirical part of the paper, Bartels does not focus on this coordination aspect (i.e., Hart versus other non-Mondale candidates), but rather analyzes the dynamic aspects of how expectations about the candidates’ winning chances influenced voters’ preferences. 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.

An exception to this is the very nice paper by Knight and Schiff (2007), who provide both a theoretical model and an empirical study of the 2004 Democratic primary. In their model, voters in different states receive signals concerning the candidates’ valences, as well as state-specific and individual preference shocks. Voters in later districts use the voting results in the earlier districts to update on the valence signal there (which, in contrast to our model, cannot be done perfectly in their model). Then, voters vote for their highest-ranked candidate, given their current beliefs. Knight and Schiff use their model to empirically estimate the degree of social learning in the 2004 Democratic primaries with a discrete choice model and find substantial momentum effects are present. Voters in early primary states have a much higher influence on the selection of the nominee than voters in later primary states.

There are two main, and interrelated, differences between Knight and Schiff (2007) and our model. First, apart from independently distributed idiosyncratic and state-level preference shocks, voters in their model care only about candidate valence. In this sense, all candidates are homogeneous. In contrast, our model formalizes the notion that some candidates are closer substitutes for each other than other candidates, and analyzes its implications.
The second main difference is that this horizontal differentiation between candidates generates welfare issues that are absent in Knight and Schiff’s model. In particular, with sufficiently many states, 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). For example, they 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. Their model is not designed to investigate what the optimal organizational form for primaries is, and thus it remains unclear why parties would choose to set up or allow to continue a sequential primary system if their only concern is to aggregate voter information about valence. In contrast, in our model, the problems of voter learning about candidate quality and vote splitting within the different “camps” interact in a way that sequential and simultaneous primary systems have both advantages and disadvantages. The central question of our paper is thus normative: Under which conditions is a sequential or a simultaneous primary system preferable from a social ex-ante point of view?

Most of the theoretical literature on primaries, reviewed below, has focused on elections with two alternatives (in which, naturally, the issue of coordination does not arise). An exception is Hummel (2008) who analyzes a model with three candidates, in which one half of the voters vote early and, after seeing those results, the other half vote. The winner is the candidate who receives the most votes overall. Voters know their preferences over candidates, but are uncertain about the preference distribution in the general population. Hummel shows that, while simultaneous elections often feature equilibria in which voters vote sincerely, this is not often the case in a sequential election system. In particular, in the late election, voters are likely to condition on the first period election result because it provides information about which candidates are likely to be the top contenders. A voter is much more likely to be pivotal between the two top contenders, which endogenously induces voters to focus on candidates who did well in the first round.

Dekel and Piccione (2000) analyze a model of sequential elections in which sophisticated voters try to aggregate their private information through voting. While, in principle, more information is available for voters in later elections, they show that every equilibrium of the simultaneous game is also an equilibrium of the sequential game, regardless of the sequence. The intuition for this result is that strategic voters know that their vote only matters if they are pivotal, and hence they behave as if they knew that all other voters are evenly divided between the two candidates. Thus, it does not matter for the election outcome which candidate is supported by the early voters.

While Dekel and Piccione (2000) show that the sequential structure does not allow voters to improve upon the information aggregation result that can be obtained with simultaneous elections, Ali and Kartik (2006) show that there are equilibria of the sequential game that do not correspond to equilibria of the simultaneous game. In particular, they construct an equilibrium in posterior-based voting in the
context of a sequential election. In this equilibrium, if other voters play history dependent strategies, then it is individually optimal for each and every voter to do so as well. The information aggregation properties of such a herding equilibrium are worse than those of the equilibrium analyzed by Dekel and Piccione (2000). In summary, with two candidates, the design of a sequential primary system appears ill-advised from a social point of view, as the expected voting outcome is either the same or worse than in a simultaneous system.

Callander (2007) studies a sequential voting model which, on top of common value preferences, voters have an exogenous desire to vote for the winning candidate. Callander obtains equilibria which, at some point in the sequential election process, display bandwagon effects with certainty because the desire to conform eventually dominates information based voting. Battaglini (2005) shows that, when voting is costly, the set of equilibria under simultaneous and sequential models are generically disjoint. In a related experimental paper, Battaglini, Morton, and Palfrey (2007) explore empirically the implications of voting costs in sequential and simultaneous elections.

Klumpp and Polborn (2006) analyze a contest model of sequential primaries in which two competing candidates choose how to allocate their campaign expenditures on the different states that hold their elections sequentially. In equilibrium, candidates allocate a large portion of their budget to the initial states. There is momentum in the sense that the currently leading candidate has an increased equilibrium probability of winning the next election. From a normative perspective, they show that a sequential organization of primaries has the advantage of leading to lower expected expenditures than a simultaneous system.

3 The model

Let $J = \{1, \ldots, J\}$ denote the set of candidates who compete for their party’s nomination, and let $j$ denote a typical candidate. The set of states (i.e., electoral districts) is $\{1, \ldots, S\}$, with typical state $s$. We assume for simplicity that the number of states, $S$, is large and that all of them have the same size so that the candidate who wins the most states wins the nomination (with ties broken by a coin flip). States vote sequentially, though some states may vote at the same time. Voters can observe the outcome in all states that voted before their own state. The set of candidates in later elections may be a strict subset of the set of candidates in early elections, as some candidates may drop out.

Candidates differ in two dimensions. First, parameter $v_j$ measures Candidate $j$’s valence (which is a characteristic like competence appreciated by all voters). Second, there is a policy issue on which candidates have either position 0 or 1. Without loss of generality, we assume that the first $j_0$ candidates are fixed at $a_j = 0$, while the other $j_1 = J - j_0$ candidates are fixed at $a_j = 1$.

The policy dimension is meant to capture the notion that some candidates are quite similar to each other and hence close (policy) substitutes for most voters, while there is a more substantial difference to some other candidates. Other issues are treated stochastically via the incorporation of a composite
preference shock, as detailed below. The assumption that policy differences can be expressed in binary form follows Krasa and Polborn (2010), and the assumption that there is only one major fixed characteristic greatly simplifies the empirical analysis.

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

$$U^i_j = v_j - \lambda|a_j - \theta^i| + \epsilon^i_j. \quad (1)$$

Here, $\theta^i$ is voter $i$'s preferred position on the policy issue, and $\lambda$ measures the weight of the policy issue relative to valence. The proportion of the total population in district $s$ with preference for $a = 1$ is $\mu^s \in (0, 1)$, which is common knowledge among all players.

The last term, $\epsilon^i_j$, drawn from $N(0, \sigma^2)$ is an individual preference shock of voter $i$ for Candidate $j$, as in probabilistic voting models.\footnote{See, e.g., Lindbeck and Weibull (1987), Coughlin (1992) or Persson and Tabellini (2000) for a review of the various developments of this literature.} One possible interpretation of this term is that candidates also differ in a large number of other dimensions for which voters have different preferences. In this interpretation, the policy dimension modeled explicitly ($a_j = 0$ or $a_j = 1$) should be understood as the most important policy dimension.

Voters are uncertain about the candidates’ valences. Specifically, each candidate’s valence is an independent draw from a normal distribution $N(0, \sigma^2)$. Voters cannot observe $v_j$ directly. Instead, voters in electoral district $s$ observe a signal $Z^s_j = v_j + \eta^s_j$, where the additional term for Candidate $j$, $\eta^s_j$, is an independent draw from a normal distribution $N(0, \sigma^2)$. Note that $\eta^s_j$ is a district-specific observation error (as opposed to a voter-specific observation error). The idea is that voters in the same state receive their news about the candidates from the same (local) news sources so that the errors (if any) are not individual-specific. Note that, if instead observation error terms were individual-specific, then the true valence of candidates would be perfectly known after the election results of the first district, which appears unrealistic.\footnote{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.}

Also, we assume that signals are state-specific rather than national, so that election results are informative for voters in later states. Even if information arrives from national news media, it appears likely that voters are particularly attentive before a state-wide election, while most voters who live in states that will only vote in a month or so may forget today’s news stories before they make up their mind about whom they should vote for. Also, information may be interpreted differently in different states, for example because voters may have different experiences with politicians adopting the same rhetoric/positions. If, instead, all information was broadcast nationally to all voters, then election results would not be incrementally informative about candidate valence.

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 $s$. Let $J^r$ be the set of
“relevant” candidates in period $t$ elections. We assume that the set $J^t$ is known to all voters, and that each voter votes sincerely given this set of relevant candidates.\footnote{In elections with more than two candidates, there are generally very 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)).} That is, voter $i$ in district $s$ (which votes at time $t$) votes for Candidate $j$ if and only if

$$j \in \arg\max_{j' \in J^t} v^s_{j'} - \lambda|a_{j'} - \theta_i^s| + \varepsilon^s_{j'}.$$ \footnote{Since the distribution of $\varepsilon$ 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.}

Thus, the set of relevant candidates captures our notion of coordination among candidates and/or voters in later primaries. In practice, there are two ways how a candidate who participated in earlier rounds of elections may drop out of the set of relevant candidates, either by being generally considered to be a lost cause by all or most voters, or by officially withdrawing from the race. It is important to stress that a sequential structure of primaries \textit{facilitates} coordination (and the particular form of coordination that we focus on is, in our opinion, fairly natural), but, of course, sequential primaries do \textit{not enforce} any particular form of coordination. We discuss this issue further below.

### 4 A tractable model illustrating coordination and learning in primaries

In this section, we focus on a particular special case of the model that can be solved in closed form and provides some intuition for the effects of the temporal organization of primaries. There are initially three candidates ($J = 3$). Candidate 1’s position is $a_1 = 0$, while candidates 2 and 3 have $a_2 = a_3 = 1$. Furthermore, we assume that $\lambda$ is sufficiently large relative to the span of the distributions of valence $v$ such that a difference in the policy dimension (almost) always dominates both valence difference and the idiosyncratic preference shock $\varepsilon$. In other words, all voters with preferred position $\theta_i^s = 0$ vote for Candidate 1, while those voters with $\theta_i^s = 1$ either vote for Candidate 2 or 3.\footnote{In principle, the distribution of $\varepsilon$ is unbounded such that there are some voters with, say, type $\theta = 1$, but a very large $\varepsilon_1$, who thus prefer Candidate 1. However, when $\lambda$ is large relative to $\sigma_\varepsilon$, such voters will be exceedingly rare, and we just ignore these cases in this section (in order to gain tractability).} This creates a coordination problem for those voters whose preferred position is 1: If candidates 2 and 3 split the votes of those 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 also assume that the proportion of the total population with preference for $a = 1$ is equal to $\mu$ in all 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.
It is useful to denote by $\phi_\alpha$, $\alpha \in \{v, \eta, \epsilon\}$, the probability density function of the normal distribution of variable $\alpha$. (Remember that all three variables are normally distributed with expected value 0, but different variances $\sigma_\alpha^2$, $\alpha \in \{v, \eta, \epsilon\}$). Furthermore, we denote by $\phi$ and $\Phi$ (without a subscript) the pdf and the cdf of the standard normal distribution $N(0, 1)$, respectively.

We assume that the number of states is large ($S \to \infty$), and analyze two 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; in this case, the set of relevant candidates at $t = 1$ is either formed by excluding the candidate with the least votes in the first state (the two-top-finisher rule) or by excluding either Candidate 2 or 3, depending on who did worse in the first state (the top-finisher-by-position rule). Note that both rules lead to the same set of relevant candidates unless Candidate 1 finishes last in the first state.

### 4.1 Equilibrium

Proposition 1, proved in part 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. Additionally, suppose that $\lambda$ is large relative to $\sigma_v$ and $\sigma_\epsilon$.

1. If $\mu < 1/2$, Candidate 1 is the Condorcet winner and wins the nomination under both a simultaneous and a sequential primary system.

2. If $1/2 < \mu < 2/3$, candidate 1 is the Condorcet loser, and the candidate with the higher valence among Candidate 2 and 3 is the Condorcet winner.
   
   In a sequential primary system, either Candidate 2 or Candidate 3 wins. The probability that the Condorcet winner wins is $\sqrt{2} \sigma_v \left[ 1 - \frac{\arctan(\frac{\sigma_3}{\sigma_v})}{\pi} \right]$, which is decreasing in $\sigma_\eta$ and increasing in $\sigma_v$.
   
   In a simultaneous primary system, either the Condorcet winner or Candidate 1 wins. Let $\kappa = \frac{\sigma_v^2}{\sqrt{2} \sigma_v (\sigma_v^2 + \sigma_\eta^2)}$. There exists $\mu^* \in (1/2, 2/3)$ such that Candidate 1 (the Condorcet loser) wins the nomination with positive probability for every $\mu < \mu^*$. The critical $\mu^*$ is decreasing in $\sigma_v$, increasing in $\sigma_\eta$, and ambiguous in $\sigma_\epsilon$.

3. If $\mu > 2/3$, Candidate 1 is the Condorcet loser, and the candidate with the higher valence among candidates 2 and 3 is the Condorcet winner.
   
   In a sequential primary system, the Condorcet loser cannot win, but the Condorcet winner wins with probability strictly smaller than 1. Moreover, the top-2-finisher rule leads to a weakly better outcome than the top-finisher-by-position rule.
   
   In a simultaneous primary system, the Condorcet winner wins with probability 1.
The first case of $\mu < 1/2$ is obvious: 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, but even coordination is insufficient for Candidate 2 or 3 to win.

In the second case where $\mu \in (1/2, 2/3)$, type 1 voters are in the majority, and thus either Candidate 2 or 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. Specifically, a sequential system is superior with respect to avoiding vote splitting and can thus prevent a victory of the Condorcet loser; however, the winning candidate may be of lower quality than his opponent 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 of them. However, since there is vote splitting, Candidate 1 may still win, even though he is the Condorcet loser.

To see these effects in more detail, consider first sequential elections. In the first district, Candidate 1 wins a percentage of $1 - \mu$ of the vote and thus either wins the first district, or finishes second. Thus, both the two-top-finisher rule and the top-finisher-by-position rule lead to the same set of relevant candidates in all later districts. Since $\mu > 1/2$, either Candidate 2 and Candidate 3 (whoever wins more votes in the first district) will win all of the districts that vote late. Thus, in a sequential organization of primaries, it is impossible that Candidate 1 wins if he is the Condorcet loser. However, because the signal of first-district voters is not perfect, the Condorcet winner is not necessarily the top-finisher among Candidate 2 and Candidate 3 in the first district. Intuitively, a higher $\sigma_\eta$ 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 to be the better one. If $\sigma_\eta$ increases, this increases the expected valence difference between the better and the worse candidate and thus increases the probability that the Condorcet winner wins. The exact probability that the Condorcet winner wins is derived in the Appendix.

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 $\varepsilon$ are sufficiently small for most voters, then almost all $\theta = 0$-voters 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 this vote-splitting between them may have the effect that Candidate 1 wins.

Proposition 1 states that the effect of $\sigma_\eta$ on the probability that the Condorcet loser wins a simultaneous primary is ambiguous. The intuitive reason for this is the following. If the true valence
difference \( v_2 - v_3 \) is very small, then a small \( \sigma_{\eta} \) implies that Candidates 2 and 3 split the votes of type 1 voters almost 50/50 in every state, so that Candidate 1 wins almost always. In this scenario, a higher \( \sigma_{\eta} \) leads to a less equal vote splitting and thus decreases Candidate 1’s winning probability. If, instead, \( v_2 - v_3 > 0 \) and the difference is large relative to \( \sigma_{\varepsilon} \), then almost all type 1 voters vote for Candidate 2 if \( \sigma_{\eta} \) is small. When \( \sigma_{\eta} \) increases, then the estimated difference between \( v_2 \) and \( v_3 \) in the beliefs of voters decreases, which leads to more vote splitting and, thus, a higher winning probability for Candidate 1. Thus, the effect of an increase in \( \sigma_{\eta} \) depends on the other parameters.

In the third case where \( \mu > 2/3 \), 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, the electorate’s preference distribution is sufficiently extreme for \( \mu > 2/3 \) to make up for any extent of vote splitting between Candidates 2 and 3. Candidate 1 cannot win a single district if \( \mu > 2/3 \).

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. Depending on the outcome of the first district, the Condorcet winner may be eliminated. Note that such an elimination is more likely under a top-finisher by position rule (which guarantees Candidate 1 a spot as a relevant candidate in the second round of elections) than under a top-finisher rule, where Candidate 1 may be eliminated in the first round.

5 Analysis of the general multi-candidate model

We now turn to an analysis of the model for a general number of candidates in both positions. The main objective is to derive theoretical foundations of voter updating about candidate valence and of vote-share determination for the empirical analysis in Section 6. In particular, we show how vote shares in the entire sequence of elections are determined given the fundamentals (candidate valences, the set of competing candidates, and voter initial beliefs) and the signals that voters observe over the course of the election.

We start with an analysis of the vote shares of candidates in district \( s \), given that the beliefs of voters in district \( s \) are given by the vector \( \hat{v}^s = (\hat{v}_1^s, \hat{v}_2^s, \ldots, \hat{v}_J^s) \). We then turn to the determination of \( \hat{v}^s \). Let \( J_0^s \) denote the set of candidates with position 0 who are running in district \( s \), and \( J_1^s \) the set of candidates with position 1 who are running in district \( s \). Beliefs about candidate valence, together with an individual’s idiosyncratic preferences, determine the candidate that he will vote for. In particular, a voter of type \( \theta \) votes for Candidate \( j \in J_0^s \) if and only if

\[
\hat{v}_j^s + \epsilon_j - \lambda d(j, \theta) \geq \max_{j'}(\hat{v}_{j'}^s + \epsilon_{j'} - \lambda d(j', \theta)),
\]

where \( d(j, \theta) \) measures the distance between Candidate \( j \) and voter type \( \theta \) (i.e., \( d = 0 \) if voter type and
candidate agree, and \( d = 1 \) when they disagree). For a given \( \varepsilon_j \), (2) is satisfied if and only if
\[
e_j < \hat{v}^s_j - \hat{v}^d_j + \varepsilon_j - \lambda (d(j, \theta) - d(j', \theta)) \text{ for all } j' \neq j.
\] (3)

First consider a voter of type \( \theta = 0 \). Since the \( \varepsilon \)'s are distributed independently, the probability that such a voter votes for Candidate \( j \) is
\[
\prod_{J \setminus \{j\}} \Phi \left( \frac{\hat{v}^s_j - \hat{v}^d_j + \varepsilon_j}{\sigma_e} \right) \prod_{J_1} \Phi \left( \frac{\lambda + \varepsilon_j + \hat{v}^s_j - \hat{v}^d_j}{\sigma_e} \right).
\] (4)

Integrating over the possible realizations of \( \varepsilon_j \) shows that the proportion of type 0 voters who vote for Candidate \( j \in J_0^s \) is
\[
\int_{-\infty}^{\infty} \prod_{J \setminus \{j\}} \Phi \left( \frac{\hat{v}^s_j - \hat{v}^d_j + \varepsilon_j}{\sigma_e} \right) \prod_{J_1} \Phi \left( \frac{\lambda + \varepsilon_j + \hat{v}^s_j - \hat{v}^d_j}{\sigma_e} \right) \cdot \phi_\varepsilon(\varepsilon_j) d\varepsilon_j.
\] (5)

Similarly, the share of type 1 voters who vote for Candidate \( j \) is
\[
\int_{-\infty}^{\infty} \prod_{J_0 \setminus \{j\}} \Phi \left( \frac{\hat{v}^s_j - \hat{v}^d_j + \varepsilon_j}{\sigma_e} \right) \prod_{J_1} \Phi \left( \frac{-\lambda + \varepsilon_j + \hat{v}^s_j - \hat{v}^d_j}{\sigma_e} \right) \cdot \phi_\varepsilon(\varepsilon_j) d\varepsilon_j.
\] (6)

The total vote share of Candidate \( j \in J_0^s \) is then given by the weighted average of (5) and (6), where the weights are \( (1 - \mu^s) \) and \( \mu^s \). In an analogous way, the total vote share of Candidate \( j \in J_1^s \) can be derived. Thus, the vote shares of candidates in state \( s \) satisfy the following equation system

\[
W_j^s = (1 - \mu^s) \int_{-\infty}^{\infty} \prod_{J_0 \setminus \{j\}} \Phi \left( \frac{\hat{v}^s_j - \hat{v}^d_j + \varepsilon_j}{\sigma_e} \right) \prod_{J_1} \Phi \left( \frac{\lambda + \varepsilon_j + \hat{v}^s_j - \hat{v}^d_j}{\sigma_e} \right) \cdot \phi_\varepsilon(\varepsilon_j) d\varepsilon_j + \\
\mu^s \int_{-\infty}^{\infty} \prod_{J_0 \setminus \{j\}} \Phi \left( \frac{\hat{v}^s_j - \hat{v}^d_j + \varepsilon_j}{\sigma_e} \right) \prod_{J_1} \Phi \left( \frac{-\lambda + \varepsilon_j + \hat{v}^s_j - \hat{v}^d_j}{\sigma_e} \right) \cdot \phi_\varepsilon(\varepsilon_j) d\varepsilon_j, \forall j \in J_0^s
\]

\[
W_j^s = (1 - \mu^s) \int_{-\infty}^{\infty} \prod_{J_0} \Phi \left( \frac{-\lambda + \hat{v}^s_j - \hat{v}^d_j + \varepsilon_j}{\sigma_e} \right) \prod_{J_1} \Phi \left( \frac{\varepsilon_j + \hat{v}^s_j - \hat{v}^d_j}{\sigma_e} \right) \cdot \phi_\varepsilon(\varepsilon_j) d\varepsilon_j + \\
\mu^s \int_{-\infty}^{\infty} \prod_{J_0} \Phi \left( \frac{\lambda + \hat{v}^s_j - \hat{v}^d_j + \varepsilon_j}{\sigma_e} \right) \prod_{J_1} \Phi \left( \frac{\varepsilon_j + \hat{v}^s_j - \hat{v}^d_j}{\sigma_e} \right) \cdot \phi_\varepsilon(\varepsilon_j) d\varepsilon_j, \forall j \in J_1^s
\] (7)

The one remaining piece required to compute the vote shares given the sequence of signals and the set of candidates competing in every state is the determination of the ex ante beliefs about candidate valences for the voters in each state. Consider the situation in the state(s) that vote first. Voters know that candidate valences are drawn from \( N(0, \sigma_v^2) \). In addition, voters in state \( s \) receive a state-specific signal \( Z_j^s \) that is normally distributed with expected value \( v_j \) (i.e., the true valence of Candidate \( j \)) and variance \( \sigma_q^2 \). Voters can now use Bayes’ rule to derive the ex-post density of the candidate’s valence, which is again the density of a normal distribution, but now with expected value

\[
\hat{v}^s_j = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_q^2} Z_j^s
\] (8)
and variance

\[(\sigma^2_{vj})^2 = \frac{\sigma^2_u \sigma^2_{vj}}{\sigma^2_u + \sigma^2_{vj}}. \tag{9}\]

For any subsequent state, if a voter has an ex-ante belief (i.e., before seeing his own state-specific signal) about candidate \(j\)'s valence that is distributed according to \(N(\hat{v}_{j0}, \sigma^2_{j0})\) and receives a state-specific signal \(Z^s_j\), the ex-post density of the candidate’s valence is again the density of a normal distribution, but now with expected value

\[\hat{v}^s_j = \frac{\sigma^2_{j0}}{\sigma^2_{j0} + \sigma^2_{\eta}} \hat{v}_{j0} + \frac{\sigma^2_{j0}}{\sigma^2_{j0} + \sigma^2_{\eta}} Z^s_j \tag{10}\]

and variance

\[(\sigma^2_{vj})^2 = \frac{\sigma^2_{j0} \sigma^2_{\eta}}{\sigma^2_{j0} + \sigma^2_{\eta}}. \tag{11}\]

For later reference, notice that the coefficient of the candidate valence signal observed by voters in state \(j\) takes the same value for all candidates. Thus, an increase in the values of all valence signals by a constant increases ex point 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.

We now turn to the calculation of \(\hat{v}_{j0}\). If voters can infer the signals observed in all prior states, then they can obtain \(v_{j0}\) (and \(\sigma^2_{j0}\)) by applying (8) and (9) to the states that vote in the first round, and (10) and (11) sequentially to all states that vote in subsequent rounds. Proposition 2 shows that this approach is indeed feasible: 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, as Corollary 1 shows, the valence signals \(Z^s_j\). This method can be applied recursively to recover the valence signals in all states that vote earlier.

**Proposition 2** Consider (7) as an equation system in \(\{\hat{v}^1_j, \hat{v}^2_j, \ldots\}\). There exists a unique vector of valence values \((0, x_2, x_3, \ldots, x_k)\) such that all solutions of (7) are of the form \((0, x_2, x_3, \ldots, x_k)+(c, c, \ldots, c), c \in \mathbb{R}\).

**Proof.** See Appendix. ■

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.

---

8The application of (10) and (11) is by round, i.e., all states voting in a particular round use values of \(v_{j0}\) and \(\sigma^2_{j0}\) as obtained from the signals up to the end of the previous round.
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 equation 10 are of the form \((0, y_2, y_3, \ldots, y_k) + (\gamma, \gamma, \ldots, \gamma)\).

Proof. This follows from the fact that equations (10) 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 \(\gamma\) is immaterial in determining voting shares and can be normalized to zero.

To recapitulate, this section shows that the vote shares of candidates in a sequence of state contests can be obtained on the basis of equations (7) – (8), and given (a) the number of candidates in each position in each state contest, (b) the valence of these candidates, (c) the signals for every candidate observed by the voters in each state, (d) the fraction of voters \(\mu_j\) in each state, \(j\), who are of political position 1, and (e) the values of four parameters: \(\sigma_v\), \(\sigma_\eta\), \(\lambda\), and \(\sigma_\epsilon\). We next turn to the description of the data and our estimation procedure.

6 Empirical analysis of the 2008 Democratic primaries

6.1 Data

Our dataset consists of the vote shares from the 2008 United States Presidential primary of the Democratic Party. The three candidates that are included in our analysis are Barack Obama, Hillary Clinton, and John Edwards, while we exclude Dennis Kucinich and other minor candidates. We consider primaries in all states except Michigan, plus the District of Columbia, yielding a total 50 state contests.

The prices on the Iowa Election Market for the 2008 Democratic nomination support this selection of candidates. For example, on December 31, 2007 (i.e., just before the first primaries), the Arrow...
security that paid $1 if Hillary Clinton won the nomination had an average price of 63.8 cents, and the prices for Edwards and Obama were 11.5 cents and 24 cents, respectively. Thus, the three candidates that we focus on each had perceived winning chances greater than 10 percent. In contrast, the average price for the “rest of field” contract (i.e., any other person being nominated) on 12/31/2007 was 1.7 cent. Thus, even though Kucinich did receive a non-trivial vote share in some states, the market prices indicate that he was never perceived as a plausible nominee by market participants. Since such “protest candidates” do not fit our theoretical framework, we exclude Kucinich and other minor candidates.

A key component of the model is that candidates are distinguished by their horizontal position. In the introduction, we have presented evidence that voters 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 clearly 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. Yet, even if the driving factor for the closer substitutability between Edwards and Obama was rather a male-female divide among voters, the implications for our estimation do not change.

For the three major candidates, we obtain the vote percentage in the primary or caucus of each state from the Federal Election Commission. This data, along with the information about the round in which each state voted, is presented in Table 3 in the Appendix. However, the vote shares do not sum up to 100 percent as they include votes for Kucinich and other minor candidates whom we dropped from our analysis, for candidates who have already withdrawn, or for “uncommitted” delegates. To ensure that vote shares representing serious votes sum up to 100% (as assumed by the model), we rescale all the vote shares accordingly for the purpose of econometric analysis.

### 6.2 Identification

Our data consists of the number of candidates who compete in each state contest, along with their political position, vote shares, and the round of each state contest in the primary run. We do not observe voter signals, the distribution of voters to political positions (\( \mu^s \)), or the candidate valence. We also do not observe the value of the parameters \( \sigma_v, \sigma_\eta, \) and \( \lambda \). Thus, we do not have all the information needed to calculate vote shares in primary contests for various configurations of state vote sequencing, as described at the end of section 5. With our data being obtained from a single primary run, it is not feasible to obtain credible estimates of \( \mu^s \); we instead posit that \( \mu^s \) is a random draw from the uniform distribution with mean equal to one half and support \( S_\mu \). The assumption that on average there is an approximately equal proportion of Democratic primary voters who support a “grass roots”

\footnote{Deltas and Polborn (2009) argue that the single most salient partition of the Democratic candidates in the recent past is 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.}
or “outsider” candidate against an “experienced” or “establishment” candidate is based on findings in Deltas and Polborn (2009). Based on the three Democratic primaries from 2000 to 2008, they show that the political positions of the candidates (i.e., “insider” or “outsider”) do not significantly affect the vote shares that the candidates receive.

Given that we do not estimate state specific values of $\mu_s$, inverting the vote shares to obtain the state signals is not a feasible strategy. Rather, we only aim to estimate (i) the standard deviation of candidate valence, denoted by $\sigma_v$; (ii) the standard deviation of state-specific information shocks, denoted by $\sigma_\eta$; (iii) the salience of the two major political positions, denoted by $\lambda$; and (iv) the support of electoral preferences for the two main political positions, denoted by $S_\mu$.

In the estimation, we consider the withdrawal of Edwards after the fifth state contest as exogenous. In other words, we take the number of competitors in each state as fixed. Given this, the four parameters listed above pin down the stochastic process that generates the vote shares of the 2008 Democratic primary run. These parameters can also be used to obtain the stochastic process of vote shares under different state voting sequences, and under different assumptions about how long the third candidate (i.e., the equivalent of Edwards in a future race) stays in the race. There is no way to infer what the outcome of the 2008 primary itself would be with each different rule, but we can predict how the distribution of outcomes (e.g., election probabilities) differs across different rules. More importantly, knowledge of these parameters is sufficient to predict the distribution of outcomes in other hypothetical races, assuming that the parameters remain stable over time. To make predictions for other races, we need to specify not only the sequencing of state contests, but also the point at which various candidates drop out of the race. We concentrate on these prospective simulations, and describe them in detail in section 7 below.

We now turn to a (somewhat informal) discussion of identification, where we consider the four parameters separately, taking the values of other parameters as given. This provides a useful intuition about the sources of identification, 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 $S_\mu$ and $\sigma_\eta$ are identified jointly from the time variation of vote share volatility. Holding the candidates fixed, the model predicts that share volatility declines over time as voter beliefs about candidates’ valence becomes more precisely concentrated on its true value. In the limit, once candidates’ valences become known, share variability will be driven solely by variability in $\mu_s$. Thus, holding other parameters constant, $S_\mu$ is identified from the limit share variability, and $\sigma_\eta$ is obtained from the rate of decline in share variability towards that limit.

The parameters $\lambda$ and $\sigma_v$ are identified jointly from the mean vote shares and how these change after Edwards withdraws. High values of $\lambda$, holding other parameters constant, imply that a higher percentage of voters whose first choice is Edwards will vote for Obama in the absence of Edwards, as high values of $\lambda$ make Clinton a worse substitute for Edwards. The value of $\sigma_v$ is identified from the share of candidates in the two political position as a function of the number of candidates in each
political position, both initially and in later election rounds. The higher the value of $\sigma_v$, the higher the expected difference in valence between the best of Obama and Edwards, and Clinton. Thus, higher values of $\sigma_v$ are associated with lower vote shares for Clinton.

As noted above, identification of any particular parameter comes from multiple sources of data variation, and the informal discussion above focuses on what is likely to be the main sources of identification. To see the interdependence of parameter estimates, consider the following example: A higher value of $\lambda$ would increase the value of $\sigma_\eta$ implied by any given observed vote share volatility of Clinton. Since Clinton would be a poorer substitute for the other candidates, any vote share variability could be rationalized by higher signal volatility. 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.

As a final note in identification, we observe that the share of Clinton is sufficient for all the above identification arguments to go through. We therefore only utilize her vote share for each of the 50 states participating in the primary. In fact, following the withdrawal of Edwards, the vote shares of Obama provide no additional information, given that vote shares add to 100 percent. For the five contests in which Edwards participated, his vote shares add some information, but this information is not needed for identification. Omitting this information yields substantial computational advantages, as we note below, with a very small loss of efficiency.

6.3 Estimation

We estimate the unknown parameters $\mu$, $\sigma_v$, $\sigma_\eta$, and $\lambda$ from the 2008 Democratic primary using the method of moments. Given that our emphasis is on obtaining plausible parameters values for the purpose of simulation rather than for model testing, we utilize four moments of the data based on the identification arguments outlined in the preceding section. This leads to exact identification.\textsuperscript{13} We also limit ourselves to using the vote shares of Hillary Clinton: as discussed in Section 6.2, the vote shares of Obama alone provide no additional information, and the vote shares of Edwards would add information for only 5 state contests and necessitate the addition of moments, leading to substantial increases in computational cost for little gain.

We next describe our estimation approach. Let the observed share of Clinton in state $s$ be $W_s^C$. Next, partition states into three groups. The first group consists of the 5 states for which there was a three way race between Clinton, Edwards, and Obama; denote this group by 3WAY, used (in the absence of any ambiguity) alternatively as a set or superscript. The second group consists of the 22 states that voted on Super Tuesday, denoted by ST. The last group consists of the 23 states that voted after Super Tuesday, denoted by pST. The union of the last two groups is denoted by the set or subscript 2WAY. The indicator variable $1_{s \in A}$ takes the value of 1 if state $s$ belongs in the group $A$ and zero otherwise.

\textsuperscript{13}Incorporating additional moments would increase efficiency, but at substantial computational cost, primarily due to the iterative procedure needed to obtain the optimal weight matrix.
Denote the sample average share of Clinton in the group of states A by \( W_A^C \).

Consider an election with two candidates located in position 1 and one candidate located in position 0, with features as described in the theory section above. The value of \( \mu^s \) for each state is a random draw from the uniform distribution with mean 0.5 and support \( S_\mu \). Valences and signals are distributed normally with means 0 and \( v_j \), and variances \( \sigma^2_v \) and \( \sigma^2_\eta \), respectively. There are five sequential contests in states \( s = 1, \ldots, 5 \), at the end of which the weakest of the two candidates in position 1 withdraws.\(^{14}\) The strongest of the two candidates in position 1 competes with the candidate in position 0 in two more rounds, one consisting of 22 states (\( s = 6, \ldots, 27 \)), and the other consisting of 23 states (\( s = 28, \ldots, 50 \)). The shares of the candidates are obtained from the equation system (7), with valence updating taking place on the basis of equations 10 and 11.\(^{15}\) However, following the first of the two rounds, we assume that valence is perfectly revealed (this is substantially true for the estimated parameter values, or it can be assumed separately that after a big round, candidate scrutiny is such that valence information is revealed perfectly).\(^{16}\)

Denote the sole candidate in position 0 as \( j = 0 \), the strongest of the two candidates in position 1 as \( j = 1a \) and the weakest of the two candidates in position 1 as \( j = 1b \). Next, define

\[
W_0(\alpha|v_0, v_{1a}, v_{1b}, \sigma_\eta, \lambda, S_\mu) = E[W_0(\alpha|v_0, v_{1a}, v_{1b}, \sigma_\eta, \lambda, S_\mu)]
\]

with the expectation taken with respect to the distributions of \( \mu^s \) and signal histories. Note that \( W_0^s \) does not depend on the values of \( \mu \) on other states (even those that voted before state \( s \)). Thus, this expectation can be obtained by integrating \( W_0^s \) over the distribution of signal histories (conditional on the vector of valences), and then integrating the result with respect to the distribution of \( \mu^s \). Observe that \( W_0(\alpha|v_0, v_{1a}, v_{1b}, \sigma_\eta, \lambda, S_\mu) \) is still a random variable at the start of a primary contest, because its value depends on the random valence draws.

The first moment in our analysis is based on the expectation of \( W_0(\alpha|v_0, v_{1a}, v_{1b}, \sigma_\eta, \lambda, S_\mu) \) for the first five states that vote sequentially, and is given by

\[
m_1(\sigma_v, \sigma_\eta, \lambda, S_\mu) = E_{W} \{ 1_{s \in W_0} W_0(\alpha|v_0, v_{1a}, v_{1b}, \sigma_\eta, \lambda, S_\mu) \},
\]

where the expectation is taken with respect to the joint distribution of valence draws.

Similarly, define the expected vote share of the sole candidate in position 0 competing against only

\(^{14}\)Our definition of “weakest” is the candidate with the lowest valence draw. This is clearly the case in the 2008 primary, as Obama is ex post widely understood to be of higher valence than Edwards. Note that for our point estimates, the candidate with the highest valence draw will also have the higher perceived valence at Super Tuesday with overwhelming probability. Since the candidate has early indication of the signals prior to the election round, the weaker of the two candidates will indeed withdraw at that stage.

\(^{15}\)For the very first state, updating takes place on the basis of equations 8 and 9.

\(^{16}\)For our estimated parameter values, the uncertainty of perceived valence immediately after Super Tuesday has dropped to only a quarter of its value in Iowa. This assumption greatly simplifies the computational aspect of the estimation, as we need not update candidate valence following Super Tuesday, and we can treat all subsequent states as voting simultaneously.
the best of the two candidates in position 1 by

\[ W_0\beta(s|v_0, v_{1a}, \sigma_\eta, \lambda, S, \mu) = E[W_0^s|v_0, v_{1a}, \sigma_\eta, \lambda, S, \mu] \] \tag{14}

with the expectation taken with respect to the distributions of \( \mu^s \) and signal histories. This is also a random variable at the start of the primary contest, since its value depends on the valence draws. Note that this expectation does not depend on the state at which the weakest of the two candidates in position 1 dropped out, but it depends on the order of voting of state \( s \).

The second moment in our analysis is based on the expectation of \( W_0\beta(s|v_0, v_{1a}, \sigma_\eta, \lambda, S, \mu) \) in the last 45 states of the contest, and is given by

\[ m_2(\sigma_\nu, \sigma_\eta, \lambda, S, \mu) = E_{v_2 \text{WAY}} \{ W_0\beta(s|v_0, v_{1a}, \sigma_\eta, \lambda, S, \mu) \} , \] \tag{15}

where the expectation is taken with respect to the joint distribution of valence draws.

The next moments are based on vote share variability. We first need to define

\[ W_0H_\beta(S\, T|v_0, v_{1a}, Z_{-ST}, \sigma_\eta, \lambda, S, \mu) = E[W_0^s|v_0, v_{1a}, Z_{-ST}, \sigma_\eta, \lambda, S, \mu, s \in ST] \] \tag{16}

where \( Z_{-ST} \) are the signal draw histories prior to the voting in states prior to Super Tuesday. This is the expected value of vote shares of candidate 0 in the group of states belong in Super Tuesday, conditional on candidate valence and signal draws prior to the voting in those states. The expectation integrates out the variability in the state voter preferences, \( \mu^s \), of the Super Tuesday states, as well as the signals received by the voters in those states.

The third moment, then, is given by

\[ m_3(\sigma_\nu, \sigma_\eta, \lambda, S, \mu) = E_{v, Z_{\mu}} \{ 1_{s \in ST}[W_0^s - W_0H_\beta(S\, T|v_0, v_{1a}, Z_{-ST}, \sigma_\eta, \lambda, S, \mu)] \} , \] \tag{17}

where the expectation is taken with respect to valence draws, signal histories, and the distribution of voter preferences.

Next define

\[ W_0H_\alpha(S\, T|v_0, v_{1a}, \lambda, S, \mu) = E[W_0^s|v_0, v_{1a}, \sigma_\eta = 0, \lambda, S, \mu, s \in pST] \] \tag{18}

This is the expected value of the vote shares of candidate 0 in the group of states that vote after Super Tuesday, conditional on candidate valence and assuming that this valence is known to the voters (recall that signal histories are not relevant for vote share determination after super Tuesday, since we assume that valence has been learned by that point). The variability in vote shares that this expectation integrates over is due to differences in voter preferences across states.

This leads to the definition of the last moment used in our analysis, which is given by

\[ m_4(\sigma_\nu, \lambda, S, \mu) = E_{v, \mu} \{ 1_{s \in pST}[W_0^s - W_0H_\beta(pST|v_0, v_{1a}, \lambda, S, \mu)] \} . \] \tag{19}
where the expectation is taken with respect to valence draws and the distribution of voter preferences.

Our estimates are based on the four by four system of equations that we obtain by setting the moments equal to their sample analogs, where the vote shares of Clinton are considered to be the realizations of the vote shares of the sole candidate in political location 0. In other words, we treat the 2008 Democratic run as representative in terms of the distribution of candidate valences, signals, and dispersion of voter preferences, of other Democratic primary runs. The system that generates the estimates can be written as

\[ m_1(\sigma_v, \sigma_\eta, \lambda, S_\mu) - \frac{1}{50} \sum_s \{1_{s \in \text{WAY}_E} W_C^s \} = 0, \]  
\[ m_2(\sigma_v, \sigma_\eta, \lambda, S_\mu) - \frac{1}{50} \sum_s \{1_{s \in \text{WAY}_E} W_C^s \} = 0, \]  
\[ m_3(\sigma_v, \sigma_\eta, \lambda, S_\mu) - \frac{1}{50} \sum_s \{1_{s \in \text{ST}_E} |W_C^s - \tilde{W}_C^s|\} = 0, \]  
and
\[ m_4(\sigma_v, \lambda, S_\mu) - \frac{1}{50} \sum_s \{1_{s \in \text{ST}_E} |W_C^s - \tilde{W}_C^s|\} = 0. \]

By substituting from the above expressions for the moment expressions, and replacing the expectation with respect to the indicator variables by the sum over the observations, the system can be written in terms of the contributions of each observation in the moment equations as

\[ \frac{1}{50} \sum_s \{1_{s \in \text{WAY}_E} E_v \{W_0 \alpha(s|v_0, v_{1a}, v_{1b}, \sigma_\eta, \lambda, S_\mu)\} - 1_{s \in \text{WAY}_E} W_C^s \} = 0, \]  
\[ \frac{1}{50} \sum_s \{1_{s \in \text{WAY}_E} E_v \{W_0 \beta(s|v_0, v_{1a}, \sigma_\eta, \lambda, S_\mu)\} - 1_{s \in \text{WAY}_E} W_C^s \} = 0, \]  
\[ \frac{1}{50} \sum_s \{1_{s \in \text{ST}_E} E_v \{W_0^s - W_0 \alpha(S\Gamma|v_0, v_{1a}, Z_{\Gamma}, \sigma_\eta, \lambda, S_\mu)\} - 1_{s \in \text{ST}_E} |W_C^s - \tilde{W}_C^s|\} = 0, \]  
and
\[ \frac{1}{50} \sum_s \{1_{s \in \text{ST}_E} E_v \{W_0^s - W_0 \beta(p\Gamma|v_0, v_{1a}, \lambda, S_\mu)\} - 1_{s \in \text{ST}_E} |W_C^s - \tilde{W}_C^s|\} = 0. \]

Given exact identification, one can find parameter values so that these four equations will be satisfied with equality. The expectations with respect to the distribution of valences and signals are obtained via Monte Carlo integration. Thus, the estimates we obtain contain some simulation error. The number of valence draws and sequences of signals was equal to 27,000, resulting in a simulation error that is less than 5 percent of the standard error (see Appendix for details).

\[ ^{17}\text{In general, it is not guaranteed that such a solution exists, but it does for this system of equations.} \]
6.4 Estimation Results

The estimation results and associated standard errors we obtained are \( \hat{\sigma}_v = 0.92 \pm 0.29 \), \( \hat{\sigma}_\eta = 2.8 \pm 1.9 \), \( \hat{\lambda} = 1.5 \pm 0.17 \), and \( S_\mu = 0.67 \pm 0.04 \). The standard errors are valid asymptotically as the number of candidates goes to infinity, clearly far from being satisfied in our sample. However, the standard errors are somewhat indicative of the relative confidence in our point estimates, with the dispersion in voter preferences being most precisely estimated (largely because it is pinned down by all 50 observations) and confidence in the variance of signals being least precisely estimated (largely because it is pinned down from behavior in the first few states).

Our interest lies primarily in the election simulation results, for which these point estimates are taken to be reasonable values to be used as inputs. However, before we proceed to the 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 our estimation results.

The point estimate of \( \sigma_v \) indicates that the better of two candidates in the same political position who differ in one standard deviation of valence will obtain \( \Phi(0.92) \approx 82\% \) of the voters who share the same political position when voters are cognizant of true valence. (Remember that the standard deviation of idiosyncratic preference shocks, \( \sigma_\epsilon \), is normalized to 1, so that \( \Phi \) is the cdf of \( \epsilon \).)

The point estimate of \( \lambda \) indicates that a candidate in position 0 who is one standard deviation better (in terms of valence) than a candidate in political position 1 will obtain \( \Phi(2.42) \approx 99\% \) of the voters in position 0 and \( \Phi(-0.58) \approx 28\% \) of the voters in position 1. Two candidates of equal valence but of different political positions get \( \Phi(1.5) \approx 93\% \) of the voters with the same political position and \( \Phi(-1.5) \approx 7\% \) of the voters with the opposite political position. Thus, political positions are very important, according to the model. With differences between two candidates of the same political position being nearly two standard deviations of valence, differences in political position are about as important in terms of voter effects as typical differences in valence between two candidates.

The point estimate of \( \sigma_\eta \) indicates that uncertainty about candidate valence swamps true differences in candidate valence for the first few states in the election. For example, suppose that one of the two candidates in the same position has a one standard deviation higher valence than his competitor. However, the chance that voters in the first district will actually perceive the better candidate as indeed better is only \( \Phi(0.92/2.8) \approx 0.629 \). Moreover, even if the better candidate receives the better signal and is thus perceived to be better, the perceived difference between candidates is likely to be quite small, and there will be substantial vote-splitting between the two candidates that occupy the same position.

In contrast, as argued above, if valence is known (which is almost the case at the final elections in a sequential primary system), then about 82% of the voters prefer the candidate with the higher valence over his competitor in the same position, and vote splitting will be minor.

More generally, 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 signals \((Z^s_j)_{s=1..N}\), expected valence is\(^{18}\)

\[
\frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2} \cdot \frac{\sum_{s=1}^N Z^s_j}{N}.
\]

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

\[
\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \left( \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2} \right)^2 \phi \left( \frac{x - v}{\sqrt{\sigma_v^2 / N}} \right) dx \phi(v/\sigma_v)dv = \frac{\sigma_v^4}{\sigma_v^2 + \sigma_\eta^2 / N},
\]

(28)

Note that this variance is always smaller than \(\sigma_v^2\), 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 to their signals is damped. Moreover, the variance of perceived valence is increasing in \(N\) and goes to \(\sigma_v^2\) in the limit of \(N \rightarrow \infty\); 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 our point estimates, (28) implies that the standard deviation of perceived valence is less than 0.3 in the first district, about 0.5 by the fifth district, and about 0.75 for district 20.

Finally, the point estimate of the support of \(\mu\) indicates that the percentage of voters in each political position can be as low as 17 percent of the electorate and as high as 83 percent of the electorate. In the typical state, in terms of deviation from the 50/50 voter partition, a third of the voters who support one political position and two-thirds will support the opposite political position. Suppose that two candidates of equal perceived valence compete in that state. Then, the candidate with the less popular position in the state will obtain \(\frac{1}{3}03 + \frac{2}{3}7 \approx 36\) percent of the votes and the candidate with the more popular position will obtain 64 percent of the votes. Vote shares are less variable than \(\mu^\sigma\) since a candidate obtains positive vote shares from voters in both political positions. Suppose instead that the candidate with the less popular position was one standard deviation better (in terms of valence) than the candidate with the more popular position. Then, in that same state the vote share of the better candidate (who has, however, the less popular position) would be \(\frac{1}{4}99 + \frac{3}{4}28 \approx 52\) percent of the electorate. Thus, the better candidate can overcome the typical electoral swing against him/her, but not by that much (however, the average difference between two randomly chosen candidates is in fact somewhat more than one standard deviation of valence).

7 Simulated effects of different institutions

We now use the empirical estimates derived in the last section to quantify the implications of different primary systems. Our basic approach is as follows: We always consider races with three candidates,

\(^{18}\)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.
two of whom share a position while the third one is in the other position. In each simulation run, we first draw candidate valences from the estimated normal distribution $N(0, 0.92^2)$. Among the candidates who share a position, this generates two candidates with different valences, whom we call $B$ (for “better”) and $W$ (for “worse”). The other, “solitary”, candidate is called $S$. We then draw state-specific signals according to $N(0, 2.8^2)$. 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.\footnote{As explained in Section 5, voters in later-voting states can essentially recover the realized state-specific signals of all states that voted before them.} We also draw aggregate position preferences in state $\mu_s$, from a uniform distribution on $[0.165, 0.835]$. Together with the distribution of individual preference shocks (normalized to be 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. (For the purpose of calculating aggregate vote shares, we assume that 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 5000 times to generate a probability distribution over outcomes. For example, we are interested in the proportion of times that $B$, $W$ and $S$ win the nomination.

Before we turn to the results, it is useful to stress a potential limitation of our simulations. When comparing different primary organizations, we keep the set of candidates and the distributions from which candidate valences and signals are drawn, fixed. In principle, the temporal setup of primaries may influence both the quality of signals and the decisions of potential primary candidates (and thus the composition of the field of candidates). With respect to signal quality, it is conceivable that, in a sequential setup, residents of early-voting states receive a better signal than voters in most other states (because candidates spend a lot of time campaigning in early states). If this is the case, our simulations will overestimate the performance of a simultaneous primary system relative to a sequential one.\footnote{Of course, if we believe that early states receive on average better quality valence signals, this could also be considered in the estimation.}

With respect to the composition of the candidate field, the following effect may arise. When the effect of vote splitting in a simultaneous primary is substantial when two candidates in one position compete with a sole candidate in the other position, there may be a considerable incentive to coordinate on one of the two candidates and force the second one out before the election even takes place. Moreover, even if no candidates drop out, voters may be able to use public opinion polls to effectively coordinate on one of the two candidates in a simultaneous election. If this is the case, our simulations will underestimate the performance of a simultaneous primary system relative to a sequential one.

While the argument concerning the endogeneity of the candidate set with respect to the temporal organization of primaries is theoretically valid, we believe that its practical significance is limited. In simultaneous party primaries (for state offices or U.S. Congress) in which no incumbent is running, there are often contests with several serious candidates who all receive substantial vote shares, and
where the winner’s vote share is often below 50 percent, indicating the potential importance of vote splitting. 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 Chicago and its suburbs, and there appears to have been considerable region-based vote-splitting. For example, 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. This example suggests that coordination facilitated by either candidates dropping out before the election or based on opinion polls cannot be taken for granted even in high-profile races.

We start by comparing the following three primary systems. The first system is a completely simultaneous primary in which all states vote at the same time. The second system is a completely sequential primary in which only one state votes at any given time. The third system is also a completely sequential primary, but, in contrast to the second system where we assume that all three candidates compete in all states, we now assume that all three candidates compete only for the first five states. Then, the candidate from the two that share a common position who is, after the fifth round of voting, perceived to be the weaker candidate (i.e., whose ex-ante valence estimate at the beginning of the sixth district is lower) drops out. The remaining two candidates compete in the remaining 45 districts. Table 1 summarizes the results.

<table>
<thead>
<tr>
<th></th>
<th>I: Simultaneous elections</th>
<th>II: Purely Sequential, no dropout</th>
<th>III: Sequential with dropout after 5 rounds</th>
</tr>
</thead>
<tbody>
<tr>
<td>S vote share</td>
<td>40.7%</td>
<td>38.7%</td>
<td>44.6%</td>
</tr>
<tr>
<td>B vote share</td>
<td>31.3%</td>
<td>41.2%</td>
<td>39.6%</td>
</tr>
<tr>
<td>W vote share</td>
<td>28.0%</td>
<td>20.0%</td>
<td>15.9%</td>
</tr>
<tr>
<td>S wins</td>
<td>98.4%</td>
<td>45.0%</td>
<td>39.0%</td>
</tr>
<tr>
<td>B wins</td>
<td>1.6%</td>
<td>48.9%</td>
<td>47.1%</td>
</tr>
<tr>
<td>W wins</td>
<td>0%</td>
<td>6.1%</td>
<td>13.9%</td>
</tr>
<tr>
<td>Exp. valence if S wins</td>
<td>0.016</td>
<td>0.519</td>
<td>0.578</td>
</tr>
<tr>
<td>... B wins</td>
<td>1.494</td>
<td>0.880</td>
<td>0.827</td>
</tr>
<tr>
<td>... W wins</td>
<td>n.a.</td>
<td>0.105</td>
<td>-0.012</td>
</tr>
<tr>
<td>S wins if CW</td>
<td>100%</td>
<td>88.5%</td>
<td>82.9%</td>
</tr>
<tr>
<td>B wins if CW</td>
<td>2.4%</td>
<td>68.5%</td>
<td>63.8%</td>
</tr>
<tr>
<td>Prob. that CW wins</td>
<td>35.2%</td>
<td>75.5%</td>
<td>70.2%</td>
</tr>
<tr>
<td>Winner’s exp. valence</td>
<td>0.039</td>
<td>0.670</td>
<td>0.613</td>
</tr>
</tbody>
</table>

Table 1: Simulation results
The first and second three lines provide the mean vote shares and winning percentages of candidates S, B and W in the different primary systems, respectively. The third three lines report the average valence of the nominee in the different primary systems, respectively. The next two lines give the winning probabilities of candidates S and B, conditional on being the Condorcet winner under full information. (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). Finally, the last two lines report the overall probability that the Condorcet winner wins, and the winner’s expected valence.

The results indicate that, from a welfare perspective, a completely sequential voting system without dropout (regime II) performs best, 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 I) does worst, with regime III intermediate, but closer to the pure sequential system.

For an intuition, consider first the simultaneous system. Candidate S wins almost all races, even though his average vote share is only 40.7%, because the two other candidates often split their votes almost evenly. As argued above, the variance in the voters’ perception of valence is small in the first district, and, in a simultaneous system, all states are effectively a “first” state (i.e., they only observe their own state-specific signal). Vote-splitting is thus a prevalent problem, and almost always prevents the two candidates with a shared position from winning. Expected valence of the election winner is thus close to zero, the ex-ante expected valence of Candidate S. Also, 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 (more than 1.5 standard deviations above the expected valence).

Now consider regime II, the purely sequential system in which all candidates stay in the race. The learning facilitated by the sequential structure has the effect that vote share shifts from W to B (while S’s vote share is just a bit lower than in regime I). As a consequence, B now wins much more often (48.9% of races). Note, however, that Candidate S still has an advantage in this system, as S still wins in many instances while not being the Condorcet winner. This is reflected in the candidates’ winning probability conditional on being Condorcet winner: While S wins over 88.5% of the races when he is the Condorcet winner, B wins only with probability 68.5% when he is the Condorcet winner.

In regime III, we assume that during the first five elections, all candidates compete. Then, the candidate from the two that share a common position who is, after the fifth round of voting, perceived to be the weaker candidate (i.e., whose ex-ante valence estimate at the beginning of the sixth district is lower) drops out. The remaining two candidates compete in the remaining 45 districts. From a positive point of view, this modification has the expected effect of reducing the winning probability of Candidate S (from 45% to 39%), as there is now less vote-splitting for most of the election sequence.

21Hence, all voters with $\epsilon_W \leq \epsilon_B$ (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.

22The 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 – since valence draws are iid, the probability that B’s valence is higher than S’s is 2/3.
Surprisingly though, Candidate B’s winning probability also decreases (from about 49% to 47%), while Candidate W’s winning probability increases from 6% to almost 14%. The reason is that the probability for a “mistake”, i.e., the better Candidate B being forced to drop out after 5 rounds, is quite substantial (approximately 30%). As a consequence, this system performs worse from a welfare point of view than the purely sequential system without dropout.

In terms of overall welfare of the election outcome, the difference between simultaneous and sequential elections is substantial. If we take expected valence as our welfare measure, the valence increase of $0.670 - 0.039 = 0.631 \approx 0.686\sigma_v$. Also, the probability that the Condorcet winner is selected as nominee is substantially higher under sequential voting than under simultaneous voting.

Regime III in Table 1 provides just one sequential voting regime with dropout. Together with Regime II (which can be interpreted as “dropout” after 50 rounds), it raises the question when the socially optimal dropout time is that would optimally trade-off between coordination and learning. To investigate this question, we perform simulations of a purely sequential contest (no two states vote at the same time) in which the candidate who is perceived to be the lowest valence among B and W withdraws after state $K$. We vary $K$ from 1 to 50, and plot the results in Figure 1.23

The results show that the electoral prospects of Candidate S are best for low and high values of $K$. When $K$ is low, Candidate S faces a single opponent for most states; thus, vote splitting is kept at a minimum. However, the opponent is often the low-valence Candidate W, as Candidate B can easily be eliminated by a few bad draws in the first couple of states. For high values of $K$, S faces two opponents for most races and vote splitting is substantial; thus, S also often emerges as the winner. Intermediate levels of $K$ (around 7 to 20) allow B to very likely dominate W, who then withdraws, and do so sufficiently early so that vote splitting is not excessive. This reduces the probability of winning for S. The electoral prospects of B more or less mirror those of S: They are low for low and high values of $K$ and highest for intermediate values of $K$. They peak at somewhat higher values of $K$ because a marginal increase in $K$ reduces the probability of win for candidate W almost throughout the range. Finally, the electoral prospects of W decline monotonically until nearly the very end.24

The socially optimal value of $K$ (using either reasonable measure of optimality) is even higher than the value of $K$ that maximizes B’s probability of winning. This is because higher values of $K$ provide better information for comparing B and S, conditional on these two candidates remaining in the race. While expected valence and the probability that the Condorcet winner emerges as nominee both decline for $K$ higher than about 30, the decline is very small. This suggests that for election contests of this type, the biggest concern is that the third candidate withdraws too soon rather than too late.

In practice, it may not be feasible to keep three candidates in the race for a very long time in a

23Values for $K = 1, 2, \ldots, 10$ and $K = 15, 20, 25, 30, 35, 40, 45, 50$ are as obtained from simulations, based on 25,000 replications. Values for remaining values of $K$ are linear interpolations.

24A small uptick at the end is driven by the fact that incremental increases in $K$ do not substantially affect the probability that it is Candidate W who withdraws (which is close to 1 anyway when $K$ is high), but the increase in $K$ increases W’s cumulative vote share since he competes in more states.
sequential primary system. After all, it is not just up to the candidates to decide when they want to give up, but also, voters may decide that only one of the two candidates in the shared position has a realistic probability of winning, and they may effectively eliminate a contender as a “serious candidate” even if he officially stays in the race.

Figure 1 suggests that it would be very desirable to organize the primary sequence in a way that all three candidates remain in the race for at least ten districts or so, as the increase in expected valence is steepest in that range and then flattens out. The reform proposal by the National Association of Secretaries of State (NASS) has a very good chance to achieve this objective: There are only two initial elections in Iowa and New Hampshire, followed by four regional contests of approximately twelve states voting simultaneously, respectively. It appears plausible that all candidates remain in the race (at least) until after the first large regional contest.

Table 2 therefore compares the NASS proposal (Regime V) with Regime IV whose structure is modeled after the existing primary system. Specifically, in Regime IV, there are 5 initial sequential elections, followed by “Supertuesday”, and another round in which all remaining states vote. Like in Regime III, the candidate perceived as weaker after the fifth election drops out.

\[\begin{array}{|l|c|c|}
\hline
& IV: 2008 primary sequence w/ dropout after 5 districts & V: NASS proposal w/ dropout after first regional primaries \\
\hline S vote share & 47.0% & 42.9% \\
B vote share & 36.7% & 42.0% \\
W vote share & 16.3% & 15.2% \\
\hline S wins & 37.5% & 38.7% \\
B wins & 45.7% & 50.8% \\
W wins & 16.8% & 10.5% \\
\hline Exp. valence if S wins & 0.458 & 0.640 \\
\ldots B wins & 0.726 & 0.808 \\
\ldots W wins & -0.173 & 0.014 \\
\hline S wins if CW & 65.9% & 81.9% \\
B wins if CW & 56.9% & 69.0% \\
\hline Prob. that CW wins & 59.9% & 73.4% \\
Winner’s exp. valence & 0.474 & 0.640 \\
\hline
\end{array}\]

Table 2: Simulation results: Status quo vs. NASS proposal

\[25\text{In reality, voting was more spread out after Supertuesday, but there are computational savings in assuming that all remaining states after Supertuesday vote simultaneously, and the disadvantage is very small, because voters’ valence estimates are already very precise after } 5 + 22 = 27 \text{ signals have been observed.}\]
From Table 2, it is apparent that the NASS structure does a considerably better job at eliminating the low valence candidate W, whose winning probability decreases from 16.8% to 10.5%. Interestingly, while most of those cases where W would win in Regime IV lead to a victory of Candidate B under the NASS structure, Candidate S’s winning probability also increases, as S, while facing a stronger opponent more often, also benefits from vote splitting in 14 rather than just 5 districts.

Unconditional expected valence, as well as all conditional expected valences increase. This is intuitive for Candidate S, as his expected opponent is now stronger and so, if S manages to win nevertheless, he must be pretty good. Also, expected valence conditional on W winning increases because winning is relatively hard for W in the NASS structure: To have a chance of winning, it must be true that W’s valence realization is very close to B’s (so that he is wrongly perceived as stronger even after 14 signals), and W’s valence must be substantially higher than S’s, because otherwise S would be able to capitalize on B and W splitting votes for 14 districts.

Finally, expected valence conditional on B winning increases. For B, winning becomes both easier and harder under the NASS proposal. A positive effect for B is that his probability of being (wrongly) eliminated in favor of candidate W decreases from 29.8% in Regime IV to 21.8% under the NASS structure. Yet, the increased vote splitting under the NASS structure means that winning conditional on not being eliminated becomes slightly harder for B, which increases B’s expected valence conditional on winning.

Our second welfare measure, the probability that the Condorcet winner is selected as nominee also increases substantially under the NASS proposal relative to the status quo, from 59.9% to 73.4%. Interestingly, this increase is driven by a relatively uniform increase in both S and B’s winning probability conditional on being the Condorcet winner.

8 Conclusion

At the beginning of presidential primaries, there are often several serious contenders. Some of them may be ideologically close substitutes for voters, while the difference to other candidates may be more significant. 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. The sequential nature of the primaries therefore likely has 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.

In this paper, we have presented a model of voting in sequential primaries based on the ideas of coordination and learning about candidate quality. We show that, from an ex-ante perspective, the coordination afforded by sequential elections may be beneficial or detrimental. While sequential elec-
tions have the advantage of allowing voters to coordinate (and thus avoid that a candidate wins just because his ideological opponents split the votes of their supporters among each other), the disadvantage of sequential elections is that, once coordination has occurred, there is no possibility to correct an error made in early elections. Moreover, our empirical results show that the probability of the wrong candidate dropping out after the first few primaries is substantial.

We show that sequential elections are likely to dominate simultaneous ones if valence differences between candidates are small; if the signal quality in early states is high; if there is a lot of vote-splitting between ideologically similar candidates; and if the ideological composition of the candidate field is asymmetric. In contrast, when valence differences are important, vote-splitting is not too important and the signal quality is bad, then a simultaneous primary system is superior.

We estimate the model using data from the 2008 Democratic primaries, 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. However, sequential institutions in which one of the candidate is forced out (in which therefore avoid the vote-splitting problem for most districts) are also not optimal, as a too early drop-out date induces a high probability that the better candidate drops out.

A current proposal by the National Association of Secretaries of State does very well from a welfare point of view in our simulations. According to this proposal, 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. Assuming that all candidates stay in the race until after the first large regional contest, there are sufficiently many early elections to be relatively confident that the strongest candidates survive, yet vote splitting is absent in three out of four large regional contests.
9 Appendix

9.1 Proofs

Proof of Proposition 1.

1. $\mu < 1/2$. The results follow directly from arguments given in the main text.

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

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

(29)

Since the arc tan is an increasing function and lies between 0 and $\pi$ (for positive arguments, such as here), it is easy to see that this probability is decreasing in $\sigma_\eta$ and increasing in $\sigma_v$.

2(b). $\mu \in (1/2, 2/3)$ and simultaneous elections. The voters in district $s$ observe signal $Z^i_j = v_j + \eta_j^i$. Using Bayes’ rule, the updated expected value of Candidate $j$’s valence is

$$\tilde{v}_j = \int_{-\infty}^\infty \frac{\phi\left(t\right)\phi\left(Z^i_j - t\right)}{\int_{-\infty}^\infty \phi\left(t'\right)\phi\left(Z^i_j - t'\right) dt'} dt = \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2} Z^i_j.$$  

(30)

If Voter $i$ in district $s$ has type $\theta = 1$, he votes for Candidate 2 if $\tilde{v}_2^i + \epsilon_2^i > \tilde{v}_3^i + \epsilon_3^i$, and for Candidate 3 otherwise. Rearranging, the percentage of type $\theta = 1$ voters who vote for Candidate 2 is equal to

$$\text{Prob}(\epsilon_3 - \epsilon_2 \leq \tilde{v}_2^i - \tilde{v}_3^i) = \text{Prob}\left(\epsilon_3 - \epsilon_2 \leq \frac{\sigma_v^2}{\sigma_v^2 + \sigma_\eta^2} [Z_2^i - Z_3^i]\right) = \Phi\left(\frac{\sigma_v^2 [Z_2^i - Z_3^i]}{\sqrt{2}\sigma_v (\sigma_v^2 + \sigma_\eta^2)}\right).$$

(31)

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

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

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

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

(32)

30
hence if

\[ \frac{2\mu - 1}{\mu} < \Phi \left( \frac{\sigma^2_n \left( Z_2^2 - Z_3^2 \right)}{\sqrt{2\sigma^2_n (\sigma^2_n + \sigma^2_\eta)}} \right) < \frac{1 - \mu}{\mu}. \]  

(33)

Denoting the inverse of the cumulative distribution of the standard normal distribution by \( \Phi^{-1} \), and letting \( \kappa = \frac{\sigma^2_n}{\sqrt{2\sigma^2_n (\sigma^2_n + \sigma^2_\eta)}} \), we can write this as

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

(34)

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 \( 2\kappa^2\sigma^2_\eta \). Thus, the percentage of districts won by Candidate 1 is given by

\[
\text{Prob} \left( \Phi^{-1} \left( \frac{2\mu - 1}{\mu} \right) - \kappa(v_2 - v_3) < \kappa(\eta_2 - \eta_3) < \Phi^{-1} \left( \frac{1 - \mu}{\mu} \right) - \kappa(v_2 - v_3) \right) = \\
\Phi \left( \frac{\Phi^{-1} \left( \frac{1 - \mu}{\mu} \right) - \kappa(v_2 - v_3)}{\sqrt{2\kappa \sigma_\eta}} \right) - \Phi \left( \frac{\Phi^{-1} \left( \frac{2\mu - 1}{\mu} \right) - \kappa(v_2 - v_3)}{\sqrt{2\kappa \sigma_\eta}} \right)
\]

(35)

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 given by

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

(36)

Candidate 1 wins the nomination if (35) is larger than (36) he wins more districts than Candidate 2, hence if

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

(37)

Note that the left hand side is decreasing in \( \mu \), while the right hand side is increasing in \( \mu \). Thus, if (37) holds for a particular level of \( \mu \), then it also holds for all smaller levels of \( \mu \) (equivalently, all higher levels of \( 1 - \mu \)). This is intuitive, since \( 1 - \mu \) is the percentage of voters who support Candidate 1. Let \( \mu^* \) denote the level of \( \mu \) such that (37) 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, (37) 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 \( \mu \), the same result holds (for any given \( v_2 \) and \( v_3 \)) for \( \mu \) sufficiently close to 1/2.

Now consider the case of \( \mu = 2/3 \), such that \( \frac{1 - \mu}{\mu} = 1/2 \). Since \( \Phi^{-1} (1/2) = 0 \), (37) is clearly violated.
Consider now the effect of changes in \( \sigma_\epsilon, \sigma_\eta \) and \( \sigma_v \) on (37). Note first that \( \kappa = \frac{\sigma^2_v}{\sqrt{2}\sigma_\epsilon (\sigma^2_\epsilon + \sigma^2_\eta)} \)
is decreasing in \( \sigma_\epsilon \) and increasing in \( \sigma_v \). Furthermore, the left hand side of (37) is decreasing in \( \kappa \) (as \( (1 - \mu)/\mu > 1/2 \), and thus \( \Phi^{-1} \left( \frac{1 - \mu}{\mu} \right) > 0 \)), while the right hand side is increasing in \( \kappa \) by the same argument. Thus, to preserve equality between the two sides of (37), an increase of \( \kappa \) needs to be balanced by a decrease of \( \mu^* \). Consequently, \( \mu^* \) decreases in \( \sigma_v \), and increases in \( \sigma_\epsilon \).

We now analyze the effect of \( \sigma_\eta \). Consider the difference of the left-hand and right-hand side of (37), and substitute for \( \kappa \) and set the expression equal to 0 (which implicitly determines the value of \( \mu^* \)); this yields

\[
Z = \Phi \left( \frac{\Phi^{-1} \left( \frac{1 - \mu}{\mu} \right) - v_2 - v_3}{\frac{\sigma^2_v}{\sigma_\epsilon (\sigma^2_\epsilon + \sigma^2_\eta)} \sqrt{2} \sigma_\eta} \right) - 2\Phi \left( \frac{\Phi^{-1} \left( \frac{1 - \mu}{\mu} \right) - v_2 - v_3}{\frac{\sigma^2_\eta}{\sigma_\epsilon (\sigma^2_\epsilon + \sigma^2_\eta)} \sqrt{2} \sigma_\eta} \right) = 0. \tag{38}
\]

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

\[
\frac{d\mu^*}{\partial \eta} = -\frac{\partial Z}{\partial \sigma_\eta}
\]
is the same as the sign of \( \frac{\partial Z}{\partial \sigma_\eta} \). We have

\[
\frac{\partial Z}{\partial \sigma_\eta} = \phi \left( \frac{\Phi^{-1} \left( \frac{1 - \mu}{\mu} \right) - v_2 - v_3}{\frac{\sigma^2_v}{\sigma_\epsilon (\sigma^2_\epsilon + \sigma^2_\eta)} \sqrt{2} \sigma_\eta} \right) + 2\phi \left( \frac{\Phi^{-1} \left( \frac{1 - \mu}{\mu} \right) - v_2 - v_3}{\frac{\sigma^2_\eta}{\sigma_\epsilon (\sigma^2_\epsilon + \sigma^2_\eta)} \sqrt{2} \sigma_\eta} \right) \times \frac{\sigma_v \Phi^{-1} \left( \frac{1 - \mu}{\mu} \right) \sigma^2_\eta - \sigma^2_v}{\sigma^2_v \sigma_\eta} \tag{39}
\]

(39) is greater than

\[
2\phi \left( \frac{\Phi^{-1} \left( \frac{1 - \mu}{\mu} \right) - v_2 - v_3}{\frac{\sigma^2_v}{\sigma_\epsilon (\sigma^2_\epsilon + \sigma^2_\eta)} \sqrt{2} \sigma_\eta} \right) \frac{\sigma_\epsilon \Phi^{-1} \left( \frac{1 - \mu}{\mu} \right) \sigma^2_\eta - \sigma^2_v}{\sigma^2_v \sigma_\eta} = \frac{v_2 - v_3}{\sqrt{2} \sigma^2_\eta}
\]

Since the term in square brackets goes to \( \frac{\sigma^2_v}{\sigma^2_\epsilon} \Phi^{-1} \left( \frac{1 - \mu}{\mu} \right) > 0 \) for \( \sigma_\eta \to \infty \), (39) is positive for \( \sigma_\eta \) sufficiently large. Thus, for \( \sigma_\eta \) sufficiently large, \( \frac{d\mu^*}{d\sigma_\eta} \) is positive. In contrast, for \( v_2 = v_3 \) and \( \sigma_\eta < \sigma_v \), (39) 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

\[
v_2 + \eta^s_1 > v_3 + \eta^s_3, \tag{40}
\]

\[
32
\]
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, consider first a top-finisher by position rule (so that the winner of the first districts and Candidate 1 are the two relevant candidates in later elections). The winner of the first district (either Candidate 2 or Candidate 3) gets a vote share $\mu$ 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 equation (29) 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.

Now suppose instead that the set of relevant candidates is defined by the top-finisher rule. This rule makes a difference if Candidate 1 places last in the first district (which is possible if $\mu > 2/3$). In this case, the two candidates with policy $a = 1$ go on to compete in the remaining districts, and the stronger candidate wins a majority of these districts with probability 1, just like under simultaneous elections. Note, however, that $\mu > 2/3$ does not guarantee that Candidate 1 places last, and if he does not, then the outcome is exactly the same as under the top-finisher-by-position rule. 

Proof of Proposition 2. Existence follows by construction: Since the vector $W^r$ is generated using the realized vector of estimated valences $\hat{(v^r_j)}_{j=1,...,k}$, a solution to (7) exists. Furthermore, it is clear that any vector of the form $(0, x_2, x_3, \ldots, x_k)$ also satisfies (7). 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 (7) shows that candidate $\bar{k}$ receives a strictly higher vote share than $W^r_{\bar{k}}$, a contradiction. Similarly, let $k$ be the candidate for whom $y_j - x_j$ is minimal. If $y_k - x_k < 0$, then substituting in the corresponding equation of (7) shows that candidate $k$ receives a strictly smaller vote share than $W^r_k$, a contradiction. But then, it must be true that $y_j = x_j$ for all $j = 2, \ldots, k$.

9.2 Estimation Algorithm Details

The estimation algorithm proceeds as follows. Consider a given set of parameter values, $\hat{\sigma}_v, \hat{\sigma}_\eta, \hat{\lambda}$, and $\hat{S}_\mu$. 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 $R$ normally distributed valence draws, $v^r_0$, with $r = 1, \ldots, R$, with mean zero and standard deviation $\hat{\sigma}_v$. These valence draws are assigned to the candidate in position 0. We next draw a set of $R$ pairs of normally distributed valence draws with the same standard deviation and mean. The highest of the two is labeled, $v^r_{1a}$, and the lowest $v^r_{1b}$, corresponding to the highest and lowest valence candidates from position 1, respectively.

We consider a primary election with seven rounds. For the first five rounds, all three candidates compete with each other. In the last two rounds, only the candidates with valence draws $v^r_0$ and $v^r_{1a}$ compete with each other. For each round, we evaluate vote shares for 25 different values of $\mu^r$, that
are equally spaced on a grid and are given by
\[ \mu^g = \frac{1-\delta_v}{2} + \frac{(g-1)\delta_v}{24}, \text{ for } g = 1, \ldots, 25. \]
These values essentially discretize the distribution of \( \mu^g \) and are used to compute expectations with respect to that distribution. For each value of \( \mu^g \) and each set of valence draws, we compute vote shares on the basis of equation system 7. Perceived valences are obtained on the basis of equations 10 and 11 (and their initial period variants) with signals drawn from the normal distribution centered around the true valence and with standard deviation \( \tilde{\sigma}_\eta \). Each set of valences gets an independent set of signal histories for each of the 25 values of \( \mu^g \) in the grid of \( \mu \). For the seventh round, perceived valences are assumed to be equal to the true valences.

This procedure returns seven matrices, each containing the vote shares of the candidate in political position 0 for each of the seven rounds. The rows of the matrix index different \( \mu^g \) draws and the columns different valence draws. Index each of the seven matrices by \( \rho = 1, \ldots, 7 \), and their typical element by \( \nu_{i,v}^{\rho} \). Note that our method of constructing vote shares for each round fixes the signal history for each value of \( \mu^g \). In other words, in each signal history and valence draw, the value of \( \mu^g \) is held fixed. Therefore, the vote share paths are not representative of the actual vote share paths, for which the value of \( \mu \) differs across states, and we cannot use any moments based on correlations or differences of vote shares across rounds. Our approach is valid for computing moments within a round since, as we pointed out in the text, vote shares in a particular state do not depend on voter preferences in preceding states, but only on the signals on the preceding states.

The value of \( m1(\tilde{\sigma}_v, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{S}_\mu) \) is computed by
\[
m1(\tilde{\sigma}_v, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{S}_\mu) = \frac{1}{5R} \sum_{i,v} \left( w_{1,i,v}^{\nu} + w_{2,i,v}^{\nu} + w_{3,i,v}^{\nu} + w_{4,i,v}^{\nu} + w_{5,i,v}^{\nu} \right) \tag{41}\]
The value of \( m2(\tilde{\sigma}_v, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{S}_\mu) \) is computed by
\[
m2(\tilde{\sigma}_v, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{S}_\mu) = \frac{1}{45R} \sum_{i,v} \left( 22 w_{6,i,v}^{\nu} + 23 w_{7,i,v}^{\nu} \right) \tag{42}\]
where the weights reflect the fact that there are 22 states in round 6 (Super Tuesday) and 23 states voting in round 7 (after Super Tuesday). The value of \( m3(\tilde{\sigma}_v, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{S}_\mu) \) is computed by
\[
m3(\tilde{\sigma}_v, \tilde{\sigma}_\eta, \tilde{\lambda}, \tilde{S}_\mu) = \frac{1}{R} \sum_{i,v} \left| w_{6,i,v}^{\nu} - \bar{w}_{6}^{\nu} \right| \tag{43}\]
where \( \bar{w}_{6}^{\nu} \) is the average vote share in round 6 for a given set of valence draws and signals observed by voters in prior rounds, with the average taken over the different values of \( \mu \) in the \( \mu \) grid and signals observed by voters in the current round. In other words, in computing this average, the candidates as perceived by voters at the start of the round are held “fixed,” but the voter preferences and signals

---

\textsuperscript{26}The distribution of \( \epsilon \) is discretized and evaluated at 70 equally spaced points between \(-3.5 \) and \( 3.5 \), and the the sum of the probabilities adjusted to sum to unity.
in round 6 vary. This mimics the vote share process during Super Tuesday. Finally, the value of $m_4(\tilde{\sigma}_v, \tilde{\lambda}, \tilde{\Sigma}_\mu)$ is computed by

$$m_4(\tilde{\sigma}_v, \tilde{\lambda}, \tilde{\Sigma}_\mu) = \frac{1}{R} \sum_{i, \gamma} \left| \tilde{w}^{i,\gamma}_{v} - \bar{w}_v^\gamma \right|$$

(44)

where $\bar{w}_v^\gamma$ is the average vote share in round 7 for a given set of valence draws with the average taken over the different values of $\mu$ in $\mu$ grid (recall that in round 7 we assume that valences are perfectly observed, which allows us to collapse all rounds following Super Tuesday into a single round; this assumption yields substantial computational savings). These moment values are used to calculate deviations the corresponding observed moment values in the data (reported in the equations 20 to 23). Parameter values are updated using the Newton-Raphson method until these deviations vanish (given exact identification, values for the four parameter values are found to exactly satisfy the four equation system). For the estimation, we use $R = 27,000$ resulting in very small sampling errors (on average about 4% of the standard error). This sampling error has been estimated by repeating the estimation for 26 different replications of the algorithm with different random seeds and $R = 3,000$. We calculated the standard deviation of the resulting estimates and used the fact that increasing the simulation draws by a factor of 9 decreases simulation error by a factor of 3. The point estimates of the large run are within two standard deviations of the average estimates of the 26 short estimation results (generally within 5 percent of the standard error of the point estimates). Thus, there appears to be negligible simulation bias at this number of replications. Estimation time of all runs is in the order of three weeks in a personal computer using GAUSS.
<table>
<thead>
<tr>
<th>State</th>
<th>round</th>
<th>Clinton</th>
<th>Edwards</th>
<th>Obama</th>
<th>State</th>
<th>round</th>
<th>Clinton</th>
<th>Edwards</th>
<th>Obama</th>
</tr>
</thead>
<tbody>
<tr>
<td>Iowa</td>
<td>1</td>
<td>29</td>
<td>30</td>
<td>38</td>
<td>Utah</td>
<td>6</td>
<td>39</td>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>New Hampshire</td>
<td>2</td>
<td>39</td>
<td>17</td>
<td>37</td>
<td>Louisiana</td>
<td>7</td>
<td>36</td>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>Nevada</td>
<td>3</td>
<td>51</td>
<td>4</td>
<td>45</td>
<td>Maine</td>
<td>7</td>
<td>40</td>
<td>0</td>
<td>59</td>
</tr>
<tr>
<td>South Carolina</td>
<td>4</td>
<td>27</td>
<td>18</td>
<td>55</td>
<td>Nebraska</td>
<td>7</td>
<td>32</td>
<td>0</td>
<td>68</td>
</tr>
<tr>
<td>Florida</td>
<td>5</td>
<td>50</td>
<td>14</td>
<td>33</td>
<td>Washington</td>
<td>7</td>
<td>31</td>
<td>0</td>
<td>68</td>
</tr>
<tr>
<td>Alabama</td>
<td>6</td>
<td>42</td>
<td>0</td>
<td>56</td>
<td>DC</td>
<td>8</td>
<td>24</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>Alaska</td>
<td>6</td>
<td>25</td>
<td>0</td>
<td>75</td>
<td>Maryland</td>
<td>8</td>
<td>36</td>
<td>0</td>
<td>61</td>
</tr>
<tr>
<td>Arizona</td>
<td>6</td>
<td>50</td>
<td>0</td>
<td>43</td>
<td>Virginia</td>
<td>8</td>
<td>35</td>
<td>0</td>
<td>64</td>
</tr>
<tr>
<td>Arkansas</td>
<td>6</td>
<td>70</td>
<td>0</td>
<td>26</td>
<td>Hawaii</td>
<td>9</td>
<td>24</td>
<td>0</td>
<td>76</td>
</tr>
<tr>
<td>California</td>
<td>6</td>
<td>52</td>
<td>0</td>
<td>43</td>
<td>Wisconsin</td>
<td>9</td>
<td>41</td>
<td>0</td>
<td>58</td>
</tr>
<tr>
<td>Colorado</td>
<td>6</td>
<td>32</td>
<td>0</td>
<td>67</td>
<td>Ohio</td>
<td>10</td>
<td>53</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>Connecticut</td>
<td>6</td>
<td>47</td>
<td>0</td>
<td>51</td>
<td>Rhode Island</td>
<td>10</td>
<td>58</td>
<td>0</td>
<td>40</td>
</tr>
<tr>
<td>Delaware</td>
<td>6</td>
<td>43</td>
<td>0</td>
<td>53</td>
<td>Texas</td>
<td>10</td>
<td>51</td>
<td>0</td>
<td>48</td>
</tr>
<tr>
<td>Georgia</td>
<td>6</td>
<td>31</td>
<td>0</td>
<td>67</td>
<td>Vermont</td>
<td>10</td>
<td>39</td>
<td>0</td>
<td>59</td>
</tr>
<tr>
<td>Idaho</td>
<td>6</td>
<td>17</td>
<td>0</td>
<td>79</td>
<td>Wyoming</td>
<td>11</td>
<td>38</td>
<td>0</td>
<td>61</td>
</tr>
<tr>
<td>Illinois</td>
<td>6</td>
<td>33</td>
<td>0</td>
<td>65</td>
<td>Mississippi</td>
<td>12</td>
<td>37</td>
<td>0</td>
<td>61</td>
</tr>
<tr>
<td>Kansas</td>
<td>6</td>
<td>26</td>
<td>0</td>
<td>74</td>
<td>Pennsylvania</td>
<td>13</td>
<td>55</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>Massachusetts</td>
<td>6</td>
<td>56</td>
<td>0</td>
<td>41</td>
<td>Indiana</td>
<td>14</td>
<td>51</td>
<td>0</td>
<td>49</td>
</tr>
<tr>
<td>Minnesota</td>
<td>6</td>
<td>32</td>
<td>0</td>
<td>66</td>
<td>North Carolina</td>
<td>14</td>
<td>42</td>
<td>0</td>
<td>56</td>
</tr>
<tr>
<td>Missouri</td>
<td>6</td>
<td>48</td>
<td>0</td>
<td>49</td>
<td>West Virginia</td>
<td>15</td>
<td>67</td>
<td>0</td>
<td>26</td>
</tr>
<tr>
<td>New Jersey</td>
<td>6</td>
<td>54</td>
<td>0</td>
<td>44</td>
<td>Kentucky</td>
<td>16</td>
<td>65</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>New Mexico</td>
<td>6</td>
<td>49</td>
<td>0</td>
<td>48</td>
<td>Oregon</td>
<td>16</td>
<td>41</td>
<td>0</td>
<td>59</td>
</tr>
<tr>
<td>New York</td>
<td>6</td>
<td>57</td>
<td>0</td>
<td>40</td>
<td>Montana</td>
<td>17</td>
<td>41</td>
<td>0</td>
<td>57</td>
</tr>
<tr>
<td>North Dakota</td>
<td>6</td>
<td>37</td>
<td>0</td>
<td>61</td>
<td>South Dakota</td>
<td>17</td>
<td>55</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>Oklahoma</td>
<td>6</td>
<td>55</td>
<td>0</td>
<td>31</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tennessee</td>
<td>6</td>
<td>54</td>
<td>0</td>
<td>41</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3: 2008 Democratic primary elections
Figure 1: Winning probabilities and expected valence for different dropout rounds $K$
References


Knight, B. and N. Schiff (2007). Momentum and social learning in presidential primaries. NBER.


