INTERNATIONAL ECONOMIC REVIEW
Vol. 60, No. 2, May 2019
DOI: 10.1111/iere.1

A STRUCTURAL MODEL OF ELECTORAL ACCOUNTABILITY*

BY S. BORAĞAN ARUOBA, ALLAN DRAZEN, AND RAZVAN VLAICU[1]

University of Maryland, U.S.A.; University of Maryland, U.S.A., NBER, U.S.A.; and CEPR, U.K.; Inter-American Development Bank, U.S.A.

This article proposes a structural approach to measuring the effects of electoral accountability. We estimate a political agency model with imperfect information in order to identify and quantify discipline and selection effects, using data on U.S. governors. We find that the possibility of reelection provides a significant incentive for incumbents to exert effort, that is, a disciplining effect. We also find a positive but weaker selection effect. According to our model, the widely used two-term regime improves voter welfare by 4.2% compared to a one-term regime, and better voter information about the effort of the governors would further increase voter welfare by up to 0.5%.

1. INTRODUCTION

In a democracy, elections are meant to make policymakers accountable for their performance. When elected officials are judged by the outcomes they produce, elections can improve policymaker performance in two key ways. They give incumbents who want to be reelected incentives to exert effort to improve outcomes, thus *disciplining* poor performance (Barro, 1973, considered this in a full information model, and Bernhardt and Ingberman, 1985, and Ferejohn, 1986, with asymmetric information about candidates). Elections also serve a *selection* function by screening out low performers (Banks and Sundaram, 1993; Fearon, 1999; Smart and Sturm, 2013; Duggan and Martinelli, 2015).[2,3]

One may then ask how effective elections are in performing these functions. From an empirical perspective, this is a question of how to measure the disciplining and selection effects of the electoral mechanism. Many papers, as discussed in the next section, have adopted a reduced-form approach to try to measure the effects of elections on policymaker performance. To identify the effects of elections they rely on variation in electoral incentives induced by term limits, either across electoral terms, for example, comparing reelection-eligible to lame-duck

*Manuscript received March 2017; revised May 2018.

[1] The authors thank Jim Alt, Tim Besley, and Shanna Rose for their assistance with data and general feedback, Ethan Kaplan, Nuno Limao, Emel Filiz Özbay, and seminar participants at University of Maryland, Paris School of Economics, LSE, Bocconi University, École Polytechnique, Northwestern University, Wallis Institute Conference, LACEA Annual Meeeting, SEA Annual Meeting, and IDB Research Department for useful comments, and Seth Wechsler, Pablo Cuba-Borda, and Camilo Morales-Jimenez for research assistance at various stages of the project. Please address correspondence to: S. Borağan Aruoba, University of Maryland, College Park, MD 20742. Phone: 301 405-3508. E-mail: *aruoba@econ.umd.edu*.

[2] There is a large empirical literature on the effect of elections on outcomes, termed political economic cycles. Brender and Drazen (2005, 2008) summarize key findings for political budget cycles. Welfare implications of opportunistic policymaker behavior are studied by Maskin and Tirole (2004), among others. Discipline and selection effects may also hold for indirectly elected policymakers, as discussed in Vlaicu and Whalley (2016).

[3] One may note that Fearon (1999) has been interpreted as arguing that the electoral mechanism cannot be used both to select over heterogeneous politicians and to sanction low effort. The argument simply put is that when voters are forward-looking, they will always vote for the candidate who is expected to deliver higher utility, making a purely retrospective voting rule to sanction poor performance nonoptimal. However, when politician type is whether they are subject to moral hazard or not, as it is in this model, then selection and sanctioning bad behavior are fully consistent. (See also Banks and Sundaram, 1993).

officials, or across electoral regimes, for example, comparing officials serving under shorter and longer term limit regimes.

In this article, we propose a *structural* approach to measuring the discipline and selection effects of elections. We set out a political agency model of electoral accountability that is predicated on the notion that voters are imperfectly informed principals using the electoral mechanism to improve the performance of elected policymakers as their agents. We then estimate the parameters of this model with adverse selection and moral hazard to quantitatively assess the importance of discipline and selection. We also perform several counterfactual exercises to study the welfare implications of allowing the possibility of reelection and of improving voter information.

Our model mimics those U.S. states where governors have a two-term limit in office, currently the most prevalent regime. In the model, governors are of two types: "good," who have intrinsic incentives to exert high effort, and "bad," who would exert high effort only in the presence of external incentives to do so, such as the possibility of another term in office. Neither the effort level chosen by governors nor their type are observable to voters. Instead, they observe incumbent performance, an outcome that depends stochastically on effort. Voters use observed performance to decide whether or not to reelect the incumbent governor. Another difference from the literature that tries to identify discipline and selection from incumbent performance is that we use a different measure of governor performance. We provide evidence that it captures more comprehensively voter welfare compared to individual policy measures or policy outcomes.

Based on our structural parameters, we estimate outcomes that would result in the absence of electoral accountability, that is, where there is no possibility of reelection. On the basis of this, we can measure how much electoral accountability improves outcomes, as well as whether improvements come mainly through discipline or through selection. This proves to be relevant since small net effects of electoral accountability in a reduced-form analysis (such as in our replication in Section 2 of a typical reduced-form analysis using our performance data) may hide fairly large and distinct discipline and selection effects. Disentangling these effects is thus crucial in addressing the issue of electoral accountability in the political agency model, a workhorse model in political economy.

The structural model also allows us to perform counterfactual experiments to assess the welfare effects of alternative settings, where governor incentives and voter information differ. Using parameters estimated from governors limited to two terms, we estimate outcomes under these alternative conditions (such as a one-term limit, varying the cost of exerting effort, or one where the voters observe an imperfect signal about governor effort), taking into account that both the voters and the governors in the economy adjust their equilibrium behavior accordingly. The assumption of invariance of structural parameters to the electoral environment is essential in avoiding the Lucas (1976) critique.

Our main findings are as follows. We find that 52% of governors are of the good type that exerts high effort independent of which term they are in. The possibility of reelection provides a significant incentive for some bad governors to exert high effort in their first term in order to increase their chances of reelection. Compared to the case with a one-term limit, allowing a second term leads 27% of bad governors to exert high effort in their first term of office, implying a 13 percentage point increase in the fraction of all governors who exert high effort in their first term. Discipline would be stronger were it not for a stochastic relation between effort and performance (high effort does not always lead to high performance), as well as an exogenous random component to election outcomes, that is, success or failure in reelection uncorrelated with performance. The two-term-limit regime leads to an increase in voter lifetime welfare of 4.2% relative to the case of a one-term limit. About two-thirds of this gain in welfare comes from the disciplining of bad governors. The remainder comes from the selection effect, that is, more good governors surviving to the second term because better first-term performance stochastically signals high effort and hence a higher probability that the governor is of the good type. The selection effect is reduced by a mimicking effect in that high first-term effort by bad

governors makes it harder for voters to identify them as such. In the absence of mimicking, discipline and pure selection effects would be roughly the same size, but mimicking reduces the latter by about 30%.

Through various counterfactuals, we reinforce our results that discipline is more important for voter welfare than selection. For example, in a two-term setup where all bad governors are disciplined in the first term (and thus there is no selection due to mimicking), welfare improves by 6.8% over the benchmark. Conversely, when there is no discipline, performance becomes a more informative signal about governor type. Welfare rises relative to the one-term limit (where neither selection nor discipline are present), but is lower than the two-term limit. This indicates that the increase in welfare from a longer term limit is due largely to the discipline effect. We then consider a version of the model where effort is at least partially observable. This leads to increased discipline, but as in the case with fully unobservable effort, this effect is mitigated by the stochastic nature of election outcomes: As the estimated election shock favors the incumbent on average, bad incumbent governors' incentives to exert high effort are reduced. Even if effort were fully observable, only 42% of bad governors would be disciplined, leading to a 0.5% increase in welfare relative to the benchmark of unobservable effort. The welfare gain is small in part due to the mitigating effect of a decline in selection as more bad governors become disciplined. If, on the other hand, the increase in transparency is accompanied by election outcomes that are less stochastic, perhaps because elections are now won and lost more on observable governor performance instead of unobserved characteristics or random factors, this would increase discipline considerably and lead to much larger welfare gains. In the extreme case where we shut down election shocks and make effort fully observable, welfare goes up by 4.8% relative to the benchmark since all bad governors choose to exert high effort.[4]

The plan of the article is as follows. In the next section we discuss the empirical literature on the effects of electoral accountability. In Section 3 we present our political agency model with a two-term limit. Section 4 describes the model's solution, the estimation methods, and the data. We then present and discuss our estimation results and their implications in Section 5. The final section presents conclusions. The Online Appendix contains technical details.

2. LITERATURE ON ESTIMATING EFFECTS OF ELECTORAL ACCOUNTABILITY

2.1. *Reduced-Form Estimation.* There have been a number of papers using reduced-form estimation to measure the effects of term limits on politician performance.[5] For example, Besley and Case (1995, 2003), Besley (2007), and Alt et al. (2011) consider various state-level performance measures for U.S. governors, List and Sturm (2006) look at environmental policy in U.S. states, and Ferraz and Finan (2011) consider fiscal corruption of Brazilian mayors. The methodology is to compare, within a jurisdiction, the performance of reelection-eligible politicians and lame-duck politicians, that is, politicians who are in their last legal term in office. These papers find statistically significant differences in outcomes. In his excellent survey of research on electoral accountability, Ashworth (2012) points out that this difference is a net effect that may reflect both discipline and selection, as these authors also recognize.

[4] The welfare gain in this case is lower than the 6.8% we report above for the case where all governors are disciplined. The difference is due to the presence of the election shocks in the former. Without election shocks all governors serve a second term and bad governors are able to play their type. With election shocks a fraction of bad governors lose reelection and are replaced by a first-term governor who always exerts effort. This leads to fewer second terms on average and increases welfare.

[5] A different approach is natural experiments, as in Dal Bó and Rossi (2011). They use two episodes in the Argentine Congress when term lengths were assigned randomly to study the relation between term lengths and politician effort. Consistent with our findings for U.S. governors, they find that longer terms induce higher legislator effort due to a longer horizon over which to capture the returns to high effort. Yet another approach is followed by Gagliarducci and Nannicini (2013), who estimate how increasing politicians' wages affects the composition of the candidate pool and the reelection incentives of those elected. Using a regression discontinuity design and Italian mayoral elections data, they find that higher wages increase performance and do so disproportionately through attracting more competent types.

Some of the above research makes further assumptions to try to disentangle the effects. For example, Besley (2007) argues that U.S. lame-duck governors are more in tune with voter preferences, as measured by interest group ideological rankings, suggesting that performance differences reflect a selection effect. List and Sturm (2006) argue that discipline effects will dominate selection effects if the fraction of voters who vote primarily on environmental issues is sufficiently small (see footnote 8 of their paper). Ferraz and Finan (2011) argue that by comparing performance of second-term mayors with that of first-term mayors who were subsequently reelected, one can control for electoral selection into the second term. Based on this, they argue that changes in corruption levels largely reflect discipline instead of selection.

A related approach is proposed by Alt et al. (2011), who argue that discipline can be measured by comparing first-term governors who are eligible to run again with those who are not (since they face different incentives but the same degree of selection), while selection over characteristics is reflected in the relative performance of term-limited incumbents in different terms (since they have been through different levels of selection but have the same incentives). Using several policy measures and policy outcomes they cannot reject the hypothesis that the discipline and selection effects are equal in magnitude.

2.2. *A Reduced-Form Analysis of Electoral Accountability.* We begin with a reduced-form analysis of our data, following the identification strategy common in the literature to compare the performance of politicians who can run again (reelection eligible) with those that cannot (lame ducks), controlling for various observable characteristics of politicians or the electorate. Differences in performance are then associated with different effects via specific identification assumptions. As an illustration, consider average (expected) performance of a governor who has a two-term limit. Average performance in the first term can be written as *baseline* + *discipline*, whereas average performance in the second and last term is *baseline* + *pure selection* − *mimicking*. Here "baseline" captures the level of performance that would be observed in the absence of electoral accountability, that is, independent of the effect of elections. Using terminology in line with our model, "discipline" reflects the increase in performance of bad governors induced by the desire to be reelected; "pure selection" shows the increase in average performance of second term governors due to a higher fraction of "good" governors being reelected; and "mimicking," the decrease in average performance of second term governors resulting from bad governors having mimicked good governors in the first term—thus increasing their probability of reelection—and then putting in low effort in their second terms. Selection as commonly used in the literature refers to what we consider pure selection minus mimicking.

If one simply computed the performance differential between reelection-eligible governors and lame-duck governors, or, equivalently, regressed gubernatorial performance on a dummy indicating whether the governor is eligible for reelection, the coefficient would simply be the difference between the performance of first-term governors and second-term governors, that is, *discipline* − *pure selection* + *mimicking*. It should be clear that this difference in performance by itself gives no information about either the absolute or the relative sizes of the three channels, information that structural estimation will allow us to identify.

Table 1 reports our replication of a typical reduced-form analysis of the data we use subsequently to estimate our model. It uses a governor's job approval ratings (JAR) as a proxy for performance, denoted by y. We discuss JAR as a performance measure in detail in Section 4.3.1. We estimate

$$y_{ist} = \mu_t + \mu_s + \gamma E_{ist} + \text{controls} + v_{ist}, \tag{1}$$

where an observation unit is a governor i in a state s in a period t where a period can be a month, a year, or a term. In Equation (1), E_{ist} is the dummy variable showing that the governor is reelection eligible and the regression also includes state and time fixed effects and controls. Here y_{ist} is the average of JAR surveys conducted in a month, a year, or a governor's entire

TABLE 1
REDUCED-FORM ANALYSIS OF ELECTORAL ACCOUNTABILITY

	(1)	(2)	(3)	(4)	(5)	(6)
	Dependent Variable = *Job Approval Rating (JAR)*					
Reelection eligible	0.71	0.63	−1.72	5.76**	5.26**	1.98
	(2.34)	(2.22)	(2.12)	(2.74)	(2.44)	(1.88)
Survey aggregation	Month	Year	Term	Month	Year	Term
Governors	All	All	All	Winners	Winners	Winners
States	32	32	32	31	31	31
Observations	2,378	357	150	1,638	357	114

NOTE: Estimation is done via ordinary least squares (OLS). The unit of observation is a governor in a U.S. state in a month (columns 1 and 4), year (columns 2 and 5), or a term (columns 3 and 6). Columns 1–3 use all governors, whereas columns 4–6 restrict the sample to those who win reelection. The sample consists of states with a two-term limit. See Section 4.3 for sample details. All regressions include year fixed effects (column 1 also includes month fixed effects) and state fixed effects, as well as governor controls (age, age squared, gender, years of education), political controls (party of governor, same party as president). Standard errors clustered at the state level. ** denotes significance at the 5% level.

term. The coefficient γ captures the average performance difference between reelection-eligible (first-term) and lame-duck (second-term) governors and will contain the combination of the three effects as explained above.

Turning to the results in Table 1, the first three columns show that when we consider all governors, then there is no significant difference between the performance of reelection-eligible governors and those that are not. When we restrict the sample to only those governors who subsequently win reelection (columns 4–6), then we get a positive coefficient that is statistically significant for the monthly and annual analysis but not the term-level analysis. That is, we find that performance is higher when governors are in their first term, but that this depends on the level of survey aggregation used.

Given the results in the first three columns, a typical reduced-form analysis would have concluded that there is no significant effect of electoral accountability on performance. Turning to the results in columns 4 and 5, these show that once we restrict the sample to governors who subsequently win their reelection bid, performance is higher for governors in their first term. Our estimates are similar to the results in Tables 4 and 7 of Ferraz and Finan (2011), who find that in a sample of mayors serving in a two-term limit regime the effect of being reelection eligible is larger for winners than for the full sample. In the winner subsample, performance differences cannot reflect selection, as the type composition is the same in the first versus the last term. Thus, the coefficient measures discipline in the winner subsample. This may be different from discipline in the full sample, however, as winners likely exert more effort on average. The structural approach we propose below will be able to separately estimate discipline, pure selection, and mimicking for the full sample.

2.3. *Structural Estimation.* There are very few papers that use a structural instead of reduced-form approach to study the effects of elections on policymaker performance and policy outcomes.[6] Sieg and Yoon (2017) present a structural model of the effects of reelection possibilities for U.S. governors (or lack thereof due to a two-term limit) on voter welfare. Politicians differ in competence and ideology, the latter measured along a single, fiscal policy dimension. Competence, which is unobserved for candidates who have not been in office, is fully revealed in a governor's first term of office. In their first term, some ideological types will moderate their fiscal

[6] Structural estimation is relatively rare in political economy. Some examples are Merlo (1997), Diermeier et al. (2003), and Strömberg (2008). Two other recent papers, Gowrisankaran et al. (2008) and DeBacker (2011), focus on the voter decision problem as a dynamic optimization problem, similar to our approach, but in their model politician actions are probabilistic and not strategic as in our model.

policy relative to their underlying preferences in order to win reelection to a second term. In the second term, the impossibility of reelection means there is no ideological moderation. Hence, these two effects—retention of competent politicians and ideological moderation—mean that the possibility of reelection (that is, the absence of term limits) will increase voter welfare. There is the possibility, however, of a negative "tenure" effect on voter welfare, for example, due to voter "fatigue" with politicians who stay multiple terms in office. A two-term limit generally reduces voter welfare compared to unlimited reelection (up to 6%), except when the tenure effect is sufficiently negative.

Although we also consider the effects of term limits in a structural model, the papers differ in a number of key respects. First, although both papers test the political agency models as applied to U.S. governors, the ways in which governors differ from one another are fundamentally dissimilar in the two models—ideology and competence in the Sieg-Yoon model versus willingness to supply effort (which could, but need not, reflect competence) in ours, where candidate ideology plays no role. In their model, competence is fully revealed in the first term, so that the voters' decision of whether or not to reelect a first-term governor concerns the trade-off between known competence and unknown underlying candidate ideology; in our model it is over unknown willingness to supply effort. Hence, our model considers both the moral hazard and adverse selection problems in electoral agency, whereas their model focuses on adverse selection abstracting from governors' effort decision. This means that the welfare findings cannot be easily compared. In term of methodology, both papers estimate a dynamic game. They use a semiparameteric approach and we use maximum likelihood. Finally, they use the same outcome measures as Besley and Case (1995), which include economic and fiscal variables, whereas our article uses a measure not previously used in this literature.

Finan and Mazzocco (2016) consider how electoral incentives affect the allocation of spending on public goods in the Brazilian federal legislature. They structurally estimate a model emphasizing the interaction among multiple representatives, as well as their decisions to run for office, paying special attention to inefficiencies due to electoral motivations and to corruption. They find that 26% of funds are misallocated relative to the social optimum and study the welfare effects of alternative electoral institutions such as approval voting. Although the article is also concerned with the effect of electoral accountability on outcomes, the mechanisms it highlights—interaction among legislators and their decisions to run for office—are quite different from ours, as is the outcome studied (efficient versus inefficient allocation of public spending). They also look at effects on legislators (and their interactions) instead of chief executives in states.

Finally, Avis et al. (2018) study the effects of random audits of Brazilian municipalities in their use of federal funds. In addition to electoral discipline and selection effects, they consider what they term a nonelectoral discipline effect, whereby the finding of corruption may lead to judicial punishment or reputation costs. They argue that there is minimal support in their data for electoral effects of audits, with the nonelectoral explaining 94% of the reduction in local corruption from the audit program. Hence, their paper not only looks at a very different measure of performance at a different level of government than ours, but also finds that for that measure at the municipal level, the mechanism by which information disciplines incumbents is overwhelmingly nonelectoral instead of electoral.

Given the differences in outcomes considered (as in Sieg and Yoon, 2017) or in mechanisms for accountability and levels of government (as in Finan and Mazzocco, 2016; Avis, Ferraz, and Finan, 2017), we view our article and these papers as complementary. We believe that each shows how structural estimation can be useful in investigating different aspects of the effect of the elections on performance and outcomes.

3. MODEL

As our benchmark model, we start with a simple political agency model with both moral hazard (unobserved politician effort) and adverse selection (unobserved politician preferences)

that can generate stochastic policy outcomes and reelection rules.[7] Subsequent versions of the model relax some of the benchmark model's assumptions. All voters are assumed to have the same information set and preferences, allowing modeling of a single representative voter. A governor may serve a maximum of two terms. After a governor's first term, voters may choose to replace her with a randomly drawn challenger. If a governor has served two terms, the election is between two randomly drawn challengers. The equilibrium concept we use is Perfect Bayesian Equilibrium, which will be defined formally below.

3.1. *Governor Types.* All governors enjoy rents of $r > 0$ in each term they are in office. A governor is one of two types, either "good" ($\theta = G$) or "bad" ($\theta = B$), where the probability that a governor is good is $\pi \equiv \mathbb{P}\{\theta = G\}$, where $0 < \pi < 1$. Governors choose the level of their effort. The cost of exerting low effort ($e = L$) is normalized to be zero. The difference between good and bad governors is in the cost they assign to exerting high effort ($e = H$). In any term of office good governors have no cost of exerting high effort, whereas bad governors have a positive utility cost c, which is expressed as a fraction of the rents r of office.[8] For ease of exposition, we define $c(e;\theta)r$ as the cost of effort level e for a governor of type θ, where $c(H;G) = c(L;G) = c(L;B) = 0$ and $c(H;B) = c$.

We assume that, like the governor's type θ, the cost c is observed by the governor but unobserved by the electorate. A bad governor draws c from a uniform distribution on the unit interval $[0, 1]$ when first elected, where c remains the same in all terms while she is in office.[9] The governor understands that her chance of winning reelection is ρ_H if she exerts high effort and ρ_L if she exerts low effort, where in equilibrium $\rho_L < \rho_H$. Different levels of effort lead to different distributions of observed performance y (as specified formally in Equations (5) below). Hence, the reelection probabilities ρ_L and ρ_H are a combination of the performance of the governor given her effort and the probability of reelection given her performance, and they will be determined in equilibrium.

3.2. *Governors' Effort Choice.* The problem of a governor of type θ is

$$\max_{e_1,e_2} [1 - c(e_1;\theta)]r + [\mathbf{1}_H\rho_H + (1 - \mathbf{1}_H)\rho_L][1 - c(e_2;\theta)]r, \tag{2}$$

where e_i is effort in term i and $\mathbf{1}_H$ is an indicator that equals 1 if $e_1 = H$ and 0 otherwise.

The actions of a good governor are trivial—she exerts high effort in the first term ($e_1 = H$) since it is costless and strictly increases her chances of reelection. Since effort is costless and she is indifferent over effort levels in the second term, we simply assume that $e_2 = H$ as well.[10]

For a bad governor it is clear that the optimal choice for the second term is $e_2 = L$ since exerting high effort in the second term is costly and has no benefit.[11] To derive e_1, note that if

[7] In Section 4.4 on identification we discuss why a model without moral hazard, that is, with only adverse selection, would not be consistent with all the findings of the article.

[8] Note that the two types and their levels of effort should not be interpreted too literally. A bad governor can be one who is rent-seeking or otherwise not "congruent" with the voters; for example, leaders may differ in their inherent degree of "other-regarding" preferences toward voters, as discussed in Drazen and Ozbay (2019). Alternatively, a bad governor can be one who is low competence (and thus finds it very costly to exert sufficient effort to produce good outcomes) or otherwise a poor fit for the executive duties of a governor.

[9] We also considered more general specifications, including a *Beta*(a, b) distribution, where the uniform distribution we use is a special case with $a = b = 1$. However, a and b were not separately identified in our estimation. Furthermore, when we assumed $U[0, a]$ and estimated $a < 1$, the estimated value approached 1. Finally, we tried the family of distributions with cumulative distribution function (CDF) $P(c < X) = X^a$ that gave an imprecise estimate of $a = 1.6$ (s.e. 0.6), and thus $a = 1$, our benchmark specification, was not rejected.

[10] If we assumed that good types like exerting high effort, that is, $c(H;G) < 0$, she would strictly prefer $e_2 = H$. This would also follow if, consistent with what we argue below about the relation between effort and expected performance, the good type preferred higher performance.

[11] In reality, good last-term performance may of course improve opportunities after the governor leaves office. The basic point however is that for bad governors the impossibility of another term reduces a key incentive to perform well,

a bad governor exerts high effort in her first term, her payoff is $(1 - c + \rho_H)r$, and if she exerts low effort, her payoff is $(1 + \rho_L)r$. In words, by exerting high effort the governor would forgo some of the first-term rent but would increase her chances of reelection, thus enjoying the rent for an extra term. She would therefore find it optimal to exert high effort if and only if

$$c < \rho_H - \rho_L. \tag{3}$$

The voter does not observe c but understands the maximization problem that governors face. He therefore can calculate the probability δ that a bad governor exerts high effort in her first term, that is, $\delta \equiv \mathbb{P}(e_1 = H|\theta = B)$. Given the assumption of a uniform distribution for c, we may then write

$$\delta = \mathbb{P}(c < \rho_H - \rho_L) = \rho_H - \rho_L. \tag{4}$$

3.3. *Voter's Problem.* The voter lives forever and prefers higher to lower y, where y is the performance of the governor in office. The voter's utility is linear in y.

We assume that this performance variable is in part influenced by the effort choice of the governor according to the rule

$$y_i|(e_i = H) \sim N(Y_H, \sigma_y^2), \tag{5a}$$

$$y_i|(e_i = L) \sim N\left(Y_L, \sigma_y^2\right), \tag{5b}$$

for term $i = 1, 2$, where $Y_H > Y_L$. These distributions are independent across terms and across governors and do not directly depend on type. Since the variance of the two distributions is the same, if the governor exerts high effort, the outcome will be drawn from a distribution that first order stochastically dominates the one with low effort.[12] Note that we also assume that the relationship between effort and performance is independent of the governor's type or the term she is in.

We further assume probabilistic voting in that the utility of the voter is affected by a shock $\varepsilon \sim N(\mu, \sigma_\varepsilon^2)$ occurring right before the election (that is, after e_1 is chosen). This "electoral" shock may reflect last-minute news about either the incumbent or the challenger, an exogenous preference for one of the candidates, or anything that affects election outcomes that is unrelated to the performance of the governor. Hence, the existence of the election shock makes elections uncertain events given the performance of incumbents. The mean of this shock, μ, jointly with other parameters captures how attractive the incumbent is relative to the challenger, independent of expected future performance. We turn to the details of what μ exactly captures in Section 4.4.

Define $W(y_1, \varepsilon)$ as the voter's life-time expected utility after observing the first-term performance of a governor and the election shock. It can be expressed recursively as

$$W(y_1, \varepsilon) = y_1 + \beta \max_{R \in \{0,1\}} \mathbb{E}\left\{R\left[y_2 + \varepsilon + \beta W\left(y_1', \varepsilon'\right)\right] + (1 - R)W\left(y_1', \varepsilon'\right)|y_1, \varepsilon\right\}, \tag{6}$$

so that they will put in less effort than good governors and perform less well, a phenomenon that we observe in the data.

[12] When we allow the $y|e$ distribution to have a variance that depends on e, this unrestricted model has a log-likelihood that is only 0.23 log-points higher, and there are no substantive changes in the results. Given the small improvement in goodness of fit, Schwarz Information Criterion chooses our restricted model.

where β is the voter's discount factor between electoral terms, and R is the decision to reelect. After observing the performance of the incumbent governor and the election shock, the voter makes his reelection choice. If he reelects the governor, he will enjoy her second term performance as well as the election shock, which shows up as an additive term to the utility of the voter. Note that ε does not affect the type or actions of the challenger that the incumbent faces. Once the incumbent's second term is over, a new governor drawn from the pool of candidates will come to office. The successor governor will deliver a first-term performance y_1' and face a reelection shock of ε', giving $W(y_1', \varepsilon')$ utility to the voter. If the voter does not reelect the incumbent, then a fresh draw from the pool of candidates occurs. It is important to note that the voter realizes that he may have arrived at this node with (y_1, ε) in one of three ways: a good governor, a bad governor who exerted high effort, or a bad governor who exerted low effort. The voter, of course, does not know which of these is the case, but has beliefs about them.

We use $\mathbb{V}$ to denote $\mathbb{E}[W(y_1', \varepsilon')]$, namely, the voter's expected lifetime utility at the beginning of a two-period term. This is a constant since none of the stochastic variables are persistent. It can be written as

$$\begin{aligned}(7)\qquad \mathbb{V} = {} & [\pi + (1-\pi)\delta]\frac{1}{\sigma_y\sigma_\varepsilon}\int\int W(y_1', \varepsilon')\,\phi\left(\frac{y_1' - Y_H}{\sigma_y}\right)\phi\left(\frac{\varepsilon' - \mu}{\sigma_\varepsilon}\right)dy_1'd\varepsilon' \\ & + (1-\pi)(1-\delta)\frac{1}{\sigma_y\sigma_\varepsilon}\int\int W(y_1', \varepsilon')\,\phi\left(\frac{y_1' - Y_L}{\sigma_y}\right)\phi\left(\frac{\varepsilon' - \mu}{\sigma_\varepsilon}\right)dy_1'd\varepsilon',\end{aligned}$$

where $\phi(\cdot)$ represents the standard normal probability density function (PDF). Equation (7) makes explicit the voter's uncertainty with respect to the type of the governor, her effort and performance in the first term, as well as the election shock that will be drawn before the election at the end of the first term. In what follows, we proceed as if $\mathbb{V}$ is a known constant, and it will be solved as a part of the equilibrium. Note further that

$$(8)\qquad \mathbb{E}(y_2|y_1) = \hat{\pi}(y_1)Y_H + [1 - \hat{\pi}(y_1)]Y_L,$$

where $\hat{\pi}(y_1) \equiv \mathbb{P}(\theta = G|y_1)$, that is, the voter's posterior probability that the incumbent is good after observing first-term performance. Using (8) we can write $W(y_1, \varepsilon)$ as

$$(9)\qquad W(y_1, \varepsilon) = y_1 + \beta \max_{R\in\{0,1\}} \left[R\left\{\hat{\pi}(y_1)Y_H + [1 - \hat{\pi}(y_1)]\,Y_L + \varepsilon + \beta\mathbb{V}\right\} + (1-R)\mathbb{V}\right].$$

3.4. *Election.* Solving the discrete choice problem in (9), we can summarize the voting rule $R(y_1, \varepsilon)$ with the following:

$$(10)\qquad R(y_1, \varepsilon) = \begin{cases} 0 & \text{if } \varepsilon \le \hat{\varepsilon}(y_1) \\ 1 & \text{if } \varepsilon > \hat{\varepsilon}(y_1), \end{cases}$$

where $\hat{\varepsilon}(y_1)$ is defined as

$$(11)\qquad \hat{\varepsilon}(y_1) = (1-\beta)\mathbb{V} - \hat{\pi}(y_1)(Y_H - Y_L) - Y_L.$$

This shows that the incumbent will win reelection if the first-term outcome y_1 is sufficiently good (so that the voter has a high posterior probability of the incumbent being good) or if the election shock ε is not too small or too negative.

The voter uses the following Bayesian updating rule to infer the type of an incumbent:

$$\hat{\pi}(y_1) = \frac{\mathbb{P}(\theta = G)p(y_1|\theta = G)}{p(y_1)} = \frac{\pi\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)}{[\pi + (1-\pi)\delta]\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) + (1-\pi)(1-\delta)\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right)}, \tag{12}$$

where δ, as defined in (4), is the voter's (correct) assessment about the probability that a bad governor will exert high effort in her first term, and $p(.)$ represents a generic density.[13]

Denoting the reelection probability conditional on first-term performance by $\psi(y_1)$, we have

$$\psi(y_1) \equiv \mathbb{P}(R = 1|y_1) = \mathbb{P}[\varepsilon > \hat{\varepsilon}(y_1)] = 1 - \Phi\left[\frac{\hat{\varepsilon}(y_1) - \mu}{\sigma_\varepsilon}\right], \tag{13}$$

where $\Phi(.)$ denotes the CDF of a standard normal random variable.

Finally, the last piece we need is the probabilities ρ_L and ρ_H that the governor was taking as given. These can be obtained by integrating $\psi(y_1)$ with respect to the performance distributions as in

$$\rho_H = \frac{1}{\sigma_y}\int \psi(y_1)\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)dy_1, \tag{14}$$

$$\rho_L = \frac{1}{\sigma_y}\int \psi(y_1)\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right)dy_1. \tag{15}$$

To summarize the events, Figure A-1 in the Online Appendix shows a game tree of the interaction between a governor and the voter. The sequence of actions and the information structure are as follows:

1. In her first term, a good governor ($\theta = G$) chooses $e_1 = H$. A bad governor ($\theta = B$) privately observes her cost c and she chooses effort $e_1 \in \{L, H\}$. As a result of this choice, first-term performance y_1 is realized.
2. The voter observes the incumbent's performance y_1 (which determines his current period utility) but not her effort e_1 or type θ. He updates the probability that the incumbent is type G using $\hat{\pi}(y_1)$.
3. An election is held between the incumbent and a randomly drawn challenger. Based on his beliefs about the type of the incumbent $\hat{\pi}(y_1)$ and the election shock ε, the voter decides whether to retain the incumbent or replace her with the challenger. If the incumbent is not reelected, then the game restarts.
4. If the incumbent is reelected, a bad incumbent chooses $e_2 = L$ and a good incumbent chooses $e_2 = H$. Based on e_2, a performance y_2 is drawn by nature giving the utility of the voter in that term.
5. At the end of the term, a new election is held between two randomly drawn candidates and the game restarts.

[13] One may note that though we do not include competence specifically in the model, low performance could reflect incompetence by a governor even though he or she exerts high effort. From the voter's point of view, however, the distinction between a bad governor (one that chooses not to exert effort) and an incompetent one is irrelevant, as long as incompetence is also permanent. Both are expected to perform poorly in the second term, as captured by $\hat{\pi}(y_1)$.

3.5. *Equilibrium.* A strategy for a governor is a choice of whether or not to exert high effort, that is, $e_i(c) \in \{H, L\}$, in each period that she is in office, $i = 1, 2$, conditional on her (privately observed) cost of effort realization c. A strategy for the voter is a choice of whether or not to reelect the incumbent, that is, $R(y_1, \varepsilon) \in \{0, 1\}$, given the observed incumbent's first-term performance y_1, and an electoral shock realization ε. The voter updates his beliefs about the incumbent's type according to $\hat{\pi}(y_1)$.

A perfect Bayesian equilibrium is a sequence of governor and voter strategies and voter beliefs such that in every period the governor maximizes her future expected payoff, given the voter's strategy, the voter maximizes his future expected payoff given the governor's strategy, and the voter's beliefs are consistent with the governor's strategy on the equilibrium path. As the environment is stationary, equilibrium outcomes will be a collection of equilibrium objects $(\rho_H, \rho_L, \delta, \mathbb{V})$, where δ is the probability that a bad governor exerts first-term effort (equivalently, the fraction of disciplined reelection-eligible bad governors), $\mathbb{V}$ is the voter's life-time discounted utility, and ρ_H, ρ_L are reelection probabilities following, respectively, high and low first-term governor effort. Formally, we have the following definition.

DEFINITION 1. The outcome of a Perfect Bayesian Equilibrium of the game between a governor and the voter is a collection of scalars $(\rho_L, \rho_H, \delta, \mathbb{V})$ where:

1. Given ρ_L and ρ_H, a bad governor's effort strategy e_1 leads to δ and indirectly to $\mathbb{V}$.
2. Given δ and $\mathbb{V}$, the voter's reelection strategy leads to ρ_L and ρ_H.

PROPOSITION 1. *The Perfect Bayesian Equilibrium defined above exists and is unique.*

PROOF. See Online Appendix A.

To understand the uniqueness result intuitively, consider first the decision of a bad governor in her first term. Her effort choice depends on the cost of high effort c relative to the increase in the reelection probability ρ. Her maximization problem (2) implies that her decision will be to put in high effort $e_1 = H$ if her cost c is no greater than $\rho_H - \rho_L$ and to put in low effort $e_1 = L$ otherwise. Hence, her decision may be described by a cutoff $c^* = \rho_H - \rho_L$, which will be unique if the difference $\rho_H - \rho_L$ (which is obviously between 0 and 1) is unique. The nature of the representative voter's problem in (9) will clearly have a unique cutoff level in y_1 for each realization of ε as well.

Since the probability of reelection $\psi(y_1)$ is monotonically increasing in first-term performance y_1 and the distribution of y_1 under high effort $e_1 = H$ first-order stochastically dominates the distribution of y_1 under low effort $e_1 = L$, the difference $\rho_H - \rho_L$ is unique, so that δ is as well. Finally, the voter's lifetime expected utility $\mathbb{V}$ will be unique, as the voter's value function $W(y_1, e)$ is fixed given δ.

PROPOSITION 2. *In equilibrium a good incumbent always exerts high effort, a bad incumbent exerts high effort if and only if* (3) *holds,; the voter reelects the incumbent according to* (10)*, and voter beliefs about the incumbent's type are given by* (12).

PROOF. Follows from the discussion above.

3.6. *Model with Effort Signal.* In this version of the model, we allow the voter to observe a noisy signal about the effort level of the governor in the first term. We denote this signal by z_1 and assume that it is symmetric and correct with probability ζ, that is,

$$\zeta \equiv \mathbb{P}\{z_1 = H | e_1 = H\} = \mathbb{P}\{z_1 = L | e_1 = L\}, \tag{16}$$

where $\frac{1}{2} \leq \zeta \leq 1$. The parameter ζ thus measures the informativeness of the signal. If $\zeta = \frac{1}{2}$, then the signal has no content, and the model is identical to the benchmark model. If $\zeta = 1$, then the signal fully reveals the incumbent's effort level, and performance is no longer an informative signal.

The signal will only be relevant in the first term because once an incumbent is reelected, the voter has no more actions that may be informed by the signal. Thus, the only point where the signal is useful is when the voter updates his prior π that the incumbent is good. The posterior is now defined by

$$\hat{\pi}(y_1, z_1) \equiv \mathbb{P}(\theta = G | y_1, z_1) = \frac{\pi p(y_1, z_1 | \theta = G)}{\pi p(y_1, z_1 | \theta = G) + (1-\pi) p(y_1, z_1 | \theta = B)} \tag{17}$$

$$= \begin{cases} \dfrac{\pi\zeta\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)}{[\pi + (1-\pi)\delta]\,\zeta\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) + (1-\pi)(1-\delta)(1-\zeta)\,\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right)} & \text{if } z_1 = H \\ \dfrac{\pi(1-\zeta)\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)}{[\pi + (1-\pi)\delta](1-\zeta)\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) + (1-\pi)(1-\delta)\,\zeta\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right)} & \text{if } z_1 = L, \end{cases}$$

which would then be used in calculating the voter's expected utility from reelecting the incumbent and hence his reelection rule. Note that $\hat{\varepsilon}(y_1, z_1)$ and $\psi(y_1, z_1)$ also have z_1 as an argument since they depend on $\hat{\pi}(y_1, z_1)$.

The incumbent understands that there will be a noisy signal about her first-term effort that will affect her chances of reelection and uses the following expected reelection probabilities in choosing her effort decision:

$$\rho_H = \frac{1}{\sigma_y} \int [\zeta\psi(y_1, H) + (1-\zeta)\psi(y_1, L)]\,\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) dy_1, \tag{18}$$

$$\rho_L = \frac{1}{\sigma_y} \int [(1-\zeta)\psi(y_1, H) + \zeta\psi(y_1, L)]\,\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right) dy_1. \tag{19}$$

Further details are presented in Online Appendix B.

4. SOLUTION, ESTIMATION, DATA, AND IDENTIFICATION

In this section we briefly discuss our strategy for solving and estimating the benchmark model. We also present our data. Details of the methods used and the details for the extension with an effort signal are presented in the Online Appendix.

4.1. *Solution.* The model has seven structural parameters: π, β, Y_H, Y_L, σ_y, μ, and σ_ε. As the definition of perfect Bayesian equilibrium shows, given the structural parameters, finding the equilibrium amounts to finding values for ρ_H, ρ_L, δ, and $\mathbb{V}$. In the process of doing so, we need to evaluate five equilibrium mappings, $\hat{\pi}(y_1)$, $\hat{\varepsilon}(y_1)$, $R(y_1, \varepsilon)$, $W(y_1, \varepsilon)$, and $\psi(y_1)$. As we demonstrate in Online Appendix C, solving for the equilibrium boils down to a nonlinear system of two equations in two unknowns, which we solve numerically. Consistent with our equilibrium uniqueness result, we are able to find a single solution to this system of equations given any set of structural parameters.

4.2. *Estimation.* We estimate the structural parameters using Maximum Likelihood. Our data set consists of a measure of performance (for each term in office) and reelection outcomes

for a set of governors. As such, the unit of observation will be a governor stint. This can be either one or two terms, depending on whether the incumbent was reelected. Given the structure of the model, we can define the likelihood function analytically. We show the detailed expressions for the likelihood function in Online Appendix C.

We estimate six structural parameters (π, Y_L, Y_H, σ_y, μ, σ_ε) and fix $\beta = 0.85$, which represents roughly a 4% annual discount rate over a four-year term.[14] Once estimates for the structural parameters are obtained, estimates for equilibrium outcomes (ρ_H, ρ_L, δ, $\mathbb{V}$) can be directly obtained using the invariance property of maximum likelihood estimation. Standard errors are computed using the White correction for heteroskedasticity for the structural parameters and the delta method for the equilibrium outcomes.

4.3. *Data Description.*

4.3.1. *Measuring governor performance.* To estimate our model, we use data for U.S. governors. The key choice we need to make is the variable that proxies for performance y in the data. In the model, y represents something that depends on governor effort, affects voters' utility directly, and is observable to voters. Given our assumption of linear utility, it can be a measure of utility as well. Since reelection decisions depend on the performance in the first term, the measure we use needs to be a good predictor of reelection outcomes as well.

Existing empirical tests of the effect of reelection on governor performance (as discussed in Section 2 above) use either economic variables, such as state unemployment rate or real income per capita growth, or fiscal variables, such as the growth in taxes, to measure governor performance. Such variables may be indicators of governor performance, but arguably governor performance reflects a larger set of variables, only some of which are quantifiable by (or even observable to) an outside observer. Corruption, which is shown to be important by Avis et al. (2017) or Finan and Mazzocco (2016) but difficult to measure, would be such a measure. See, for example, the discussion of Alabama governor Guy Hunt in Online Appendix E. We therefore want a broader measure that might possibly capture the multifaceted nature of performance.[15] Nevertheless, for completeness's sake, in Section 5.3, as a part of our robustness checks, we show our estimation results for two commonly used economic variables.

Theoretically, governor performance in this broader sense could be captured by expert evaluations (analogous to evaluations of U.S. presidents by historians), but such ratings are scarce. We chose to use JAR of governors from surveys of voters taken at various points during a governor's term(s). A large fraction of the JAR data come from Beyle et al. (2002), and we update their data set through the end of 2014 using various online resources. Potential voters are asked to rate the governor as "excellent," "good," "fair," or "poor" or to say that they are "undecided." As a measure of performance, we calculate for each governor the fraction of respondents who classify the governor as excellent or good out of those who express an opinion, eliminating the undecided respondents.[16] We explain more precisely below how we convert this measure based on a survey taken at a specific point into a performance measure over the governor's term. The mapping from effort to observed performance as measured by JAR implicitly includes the policy choices that governors make and outcomes across different areas, where these are aggregated into a single measure. Using specific policy choices (or other

[14] When we try alternative values of β, the estimates of structural parameters do not change and only the equilibrium object $\mathbb{V}$ adjusts.

[15] Incidentally, the JAR-based performance measure is a better predictor of election outcomes than individual economic variables. In simple probit regressions, real income per capita growth and state unemployment rate have some limited success in predicting reelection outcomes. For example, state unemployment rate has a significant coefficient and the probit regression has a McFadden R^2 of 0.05. But once JAR is included in the same regression, unemployment rate loses its significance and the McFadden R^2 almost quadruples to 0.19.

[16] It is also important to point out that JAR is not a relative rating, based on a comparison with a challenger, but it is an absolute evaluation of the governor's performance in office, because the vast majority of JAR surveys are taken long before a challenger is identified. In our model, the challenger's qualities apart from his type enter through the election shock.

TABLE 2
DETERMINANTS OF JOB APPROVAL RATINGS

	(1)	(2)	(3)	(4)
Independent Variables	Dependent Variable = *JAR*			
Unemployment	−2.93***	−3.09***	−2.80***	−2.94***
	(0.70)	(0.66)	(0.77)	(0.73)
Income growth	0.72***	0.59**	0.75***	0.66***
	(0.27)	(0.24)	(0.25)	(0.24)
Population growth	1.45	1.40*	1.56	1.57*
	(0.90)	(0.82)	(0.96)	(0.85)
Age		−0.28**		−0.28**
		(0.11)		(0.11)
Lawyer		2.58		2.37
		(1.86)		(1.81)
Male		−0.83		0.27
		(3.75)		(3.42)
Served in military		−1.66		−1.77
		(2.02)		(2.00)
Years of education		−0.02		−0.02
		(0.29)		(0.28)
Population			0.75	1.00*
			(0.46)	(0.52)
Same party president			−2.16*	−2.17**
			(1.24)	(1.23)
Partisan fit			−2.28**	−1.96**
			(0.89)	(0.92)
% Voters Governor party			−0.20	−0.26
			(0.18)	(0.17)
% Voters other party			0.13	0.08
			(0.18)	(0.18)
R^2	0.39	0.41	0.42	0.44

NOTE: The regressions include state and year fixed effects. The sample consists of 1,020 governor-year observations over the period 1976–2010 and across 50 U.S. states. Second block of variables are characteristics of the governor. "Same party president" shows if the governor is from the same party as the U.S. president. See the text for the definition of "Partisan fit." "% Voters" variables show the voters affiliated with the governor's party and the other party, respectively. *, **, and *** represent significance at 10%, 5%, and 1% levels, respectively, using standard errors that are clustered at the state level.

endogenous factors) more directly as a measure of what distinguishes good from bad governors poses an arguably insurmountable problem when considering a large number of governors, namely how a number of specific measures could be combined by an outside observer into a parsimonious measure of governor performance.

The key question is then whether such approval ratings—and thus the resulting performance indicator—are a good measure of actual governor performance. There are two basic aspects of this question. First, does the JAR measure described above capture things believed to reflect true performance, such as economic variables that other studies used? Second, to what extent is this measure contaminated by things that do not reflect performance due to governor effort, such as partisan biases of survey respondents or pandering to voters?

To address these questions we regress our JAR-based performance measure on three sets of variables, and the results are reported in Table 2. The first set contains measures of state economic performance, such as state unemployment rate, growth of state per capita personal income, and state population growth.[17] The second set of variables is governor characteristics: age, gender, years of education, and whether the governor is a lawyer or served in the military. The third set is variables measuring partisanship in the state: state's population (possibly

[17] We did not include fiscal outcomes sometimes used in some of the earlier literature (for example, higher government spending) in this regression because they are viewed differently by different groups of voters.

capturing homogeneity of preferences in smaller states), whether the governor is of the same party as the U.S. president, "partisan fit," which shows a match between the party of the governor and how the state voted in the previous presidential election and the percentage of voters of the governor's and the opposing party.[18] The regressions also include state and year fixed effects.

The first column in Table 2 uses only measures of state economic performance. The second and third columns add each of the other sets of variables, one at a time, and the fourth column uses all variables. A number of things become clear. First, as the first three lines of the table across all columns make clear, the JAR-based performance measure is highly correlated in the correct direction with measures of state economic performance used in other studies. Looking at the R^2's reported in the last row, a large fraction of the explanatory power in these regressions comes from the first three variables, in addition to the fixed effects. Even without fixed effects the R^2 of the regression that includes only the first three variables (not shown) is 0.15, though in this case only unemployment rate is significant. Hence our JAR-based performance measure is indeed capturing governor performance in terms of the macroeconomic performance of the state.

Second, although most of the partisanship variables have no statistically significant effect on our governor performance measure, the measure is significantly related to both whether the governor is of the same party as the president and partisan fit. Interestingly, both variables have a *negative* effect on our JAR-based measure; that is, congruence of the governor's party affiliation with either the president's or with the voter's preference in the most recent presidential election *lowers* the JAR rating. That is, partisan effects go in the *opposite* direction of naive conventional wisdom, where "partisan bias" is that a party's adherents overrate a governor from the same party and underrate one from the opposing party. Jacobson (2006), who found an analogous negative correlation between a governor's approval rating and his or her party being in the majority, explained it as such governors needing considerable cross-party appeal to win office. Similarly, we argue that the sign of the coefficients may be reflecting a performance effect—the bar for a Democrat governor, for example, to succeed in a Republican state is higher and thus she performs better; or alternatively, more "good" Democrat governors run for election in Republican states than "bad" Democrat governors. Thus, we think the case for arguing that JAR contains a partisan bias is weak, at best. Nevertheless, in Section 5.3 we consider an "adjusted" JAR measure where we strip our benchmark JAR measure from the effects of the two partisanship variables that are significant in Table 2.

Finally, among governor characteristics, only age has a significant (negative) effect on surveyed performance. This is consistent with JAR surveys being a good measure of governor performance to the extent that age (or any trait) is correlated with effort or with the relation between effort and performance. In any case, the coefficient is small: It shows that a one standard deviation difference (eight years) in age between two otherwise identical governors is associated with a JAR difference of 2.2 points.

One final question is whether a JAR-based performance measure as a proxy for y might reflect pandering to voters. Note first that if actions seen as pandering entered job approval ratings, our measure would still satisfy the conditions listed in the first paragraph as a proxy for y and hence would allow us to estimate the model and measure the effect of reelection incentives on incumbent governor behavior. More specifically, the things the governor does that can be considered pandering would still take effort, so that in order to pander and increase their chance of reelection, bad governors might choose to exert effort. However, the link between JAR and actual (not perceived) voter welfare would be weaker in this case, and welfare statements would be problematic. On a conceptual level, arguing that JAR is so dominated by pandering and that it contains little information about a governor's true performance is essentially arguing that voters (or individuals more generally) are simply unable to assess their own well-being. Though

[18] The "partisan fit" variable follows Jacobson (2006), where it is 1 if governor's party's presidential candidate got more than 52% vote share in the state, it is −1 if governor's party's presidential candidate got less than 48% vote share in the state, and 0 otherwise.

some take this extreme position, our comparison of the JAR data to narrative accounts of governor performance suggests that JAR does reflect a reasonably accurate assessment of voter welfare.[19] Some of these narratives are presented in Online Appendix E. On the more specific question of whether JAR is dominated by pandering, it may be argued that pandering is most likely in the election year. Our results are robust to dropping JAR surveys taken in the election year, as we discuss in Section 5.4.[20]

To summarize the discussion so far, we think JAR captures what most people would consider the performance of governors. Through our regression analysis, we were able to show that at least some key macroeconomic indicators significantly influence JAR. However, JAR is much more than just one or two indicators. The R^2 of the regression in Table 2, even after many controls including state and time fixed effects, is 0.44, which means that this regression does not capture more than half of what determines JAR. Given the sensitivity checks we have done, we are confident that it is the multifaceted performance measure we need for our analysis. JAR clearly measures factors likely reflecting performance not captured by alternative univariate measures and does not show measurable evidence of being contaminated by nonperformance-related factors.

Finally, we explain how we convert individual JAR survey results to measures that cover terms of governors. To eliminate effects of the governor's reelection campaign, we use JAR up to and including June of the final year of the incumbent's first term, that is, the election year, given that U.S. gubernatorial elections take place in November. We do not restrict the second-term JAR-based measure. We take the simple average of the JAR numbers—that is, the fraction of those describing the governor's performance as "excellent" or "good," as described above—averaged over the entire term and use them as y_1 and y_2. From here on we use "JAR" to refer to the measure described in this paragraph. We revisit some of the choices we make in this section and consider alternatives in Section 5.4.

4.3.2. *Governor stints.* Our model places some important constraints on the types of governor stints we may use in the estimation. We start with the universe of all governors that served from 1950 to the present, where we have collected basic information about the governor, some of which comes from Besley (2007), including the outcomes of their reelection bids.[21] We then apply the following filters to eliminate governors who do not fit our model of a limit of two terms of equal length across governors: (i) drop governors who did not have any term limits or had a one-term limit or a three-term limit;[22] (ii) drop governors during whose stints state election laws regarding term limits changed; (iii) drop governor stints (not just the terms) where the governor was appointed, completed someone else's term, or was elected through a special election outside the state's regular electoral cycle; and (iv) drop governors who did not complete at least three years of their first term or at least two years of their second term (for example due to resignation, passing away, or being recalled).

These filters yield 169 governor stints.[23] Matching these with the JAR data we compiled, a total of 5,549 surveys, yields 93 governor stints. Due to data availability and/or absence of term

[19] There is some evidence that events not under a governor's control happening right before elections do affect voters at the margin (Wolfers, 2007; Healy et al., 2010), but this is consistent with our modeling of the election shock ε.

[20] The view that voter assessments of government performance are not dominated by pandering is consistent with the findings of Brender and Drazen (2008) in a large cross-country sample at the national level, who find that voters punish instead of reward fiscal manipulation in election years.

[21] We consider any governor that is eligible for reelection as having run for reelection; that is, we consider the choice of not running as losing. This is justified by our review of such cases where a reasonable interpretation of the events suggests that the governor decided that he or she would not be able to win reelection and either resigned or sought other alternatives. Perhaps not surprisingly, many of these governors perform quite badly in their first term, which results in being predicted by the model as "bad" governors who did not exert effort (see Table A-1).

[22] We use data on one-term governors to validate our model in Section 5.1.2.

[23] A handful of governors serve multiple stints by being elected after some period following a completed term-limited stint. We treat each stint as a separate governor. Eliminating these governors from our sample does not change our results.

TABLE 3
KEY MOMENTS—DATA AND MODEL

	Data	Model
$\mathbb{E}(y_1)$	55.32	56.71
$\mathbb{E}(y_1\|R=0)$	48.08	50.78
$\mathbb{E}(y_1\|R=1)$	59.89	60.21
$std(y_1)$	14.47	13.93
$skew(y_1)$	−0.24	−0.22
$\mathbb{E}(y_2\|R=1)$	57.91	55.64
$std(y_2\|R=1)$	13.51	14.14
$skew(y_2\|R=1)$	−0.27	−0.14
$corr(y_1, y_2\|R=1)$	0.36	0.34
$\mathbb{E}(y_2 - y_1\|R=1)$	−1.98	−4.57
$skew(y_2 - y_1\|R=1)$	−1.77	−0.19
$corr(y_1 - y_2, y_2\|R=1)$	−0.59	−0.64
$\mathbb{P}$ (R=1)	61.3%	62.9%
$\mathbb{P}(R=1\|y_1 < Y_L)$	30.0%	34.9%
$\mathbb{P}(R=0\|y_1 > Y_H)$	15.4%	21.9%

NOTE: *std*, *skew*, and *corr* stand for standard deviation, skewness, and correlation. All moments that involve y_2 are conditional on the governor winning reelection.

limits early on in our sample, except for one governor from the 1960s, our data cover governor terms from 1978 to 2014. There are 26 election years from 32 states in our sample. The average age of a governor is 56, with 19 years of education on average; 91% of the governors in our sample are male, 55% of them are from the Democratic Party, 39% have served in the military, and 46% of them are lawyers. Comparing these numbers with the population of all governors over this period, there does not seem to be a major bias in our sample.[24] We provide the basic data that we use for estimation, namely, (y_1, R, y_2), in Table A-1 in the Online Appendix.

4.4. *Identification.* In this section we discuss the identification of the structural parameters of our model as well as some of the equilibrium outcomes. We do so by pointing out some key moments in the data that help with identification. These moments are reported in Table 3. We discuss how our model matches these moments in Section 5.1.2.

First, our model assumes that there are two types of governors ($0 < \pi < 1$). Consider the possibility that there was only one type of governor. If $\pi = 0$, then there are only bad governors, whereas if $\pi = 1$, there are only good governors. In the latter case, since all governors are good, all governors would exert high effort in both terms, which means $corr(y_1, y_2) = 0$ as each term's performance for a reelected governor is drawn independently from the same distribution. On the other hand, if all governors were bad, then $corr(y_1, y_2)$ would be zero or even negative.[25] Table 3 shows that in our data, this correlation is 0.36 with a standard deviation of 0.14 (significant at the 5% level), which indicates the presence of two types.[26]

[24] Our largest sample of 588 governor stints, which includes governors we dropped with the filters above, shows 96% of governors as male, 55% were from the Democratic Party, 53% served in the military, and 54% are lawyers. The average age of a governor is 53, and they have 19 years of education on average.

[25] It is important to note that the cases where $\pi = 0$ and $\pi = 1$ are not covered by our model since updating of beliefs about types as in (12) cannot occur. When $\pi = 1$ the reelection rule does not matter as all governors are good and they always exert effort. When $\pi = 0$, one can consider various other election schemes, under some of which, we may have some discipline. If this is the case, then for these disciplined governors, y_1 will be drawn from the high-effort distribution and y_2 will be drawn from the low-effort distribution, which creates a negative correlation. Alternatively, if there are no disciplined bad governors, then to the extent any governor is reelected, $corr(y_1, y_2) = 0$ for the same reason as we explained for good governors.

[26] We also assume that there are two levels of effort and not more. One may think, for example, that if given the choice good governors may choose to exert a very high level of effort to signal their type. If this was the case in the data, we would expect a positive correlation between $y_1 - y_2$ and y_2, where the latter proxies for the type of the governors.

TABLE 4
PARAMETER ESTIMATES

Structural Parameters						Equilibrium Objects			
π	Y_L	Y_H	σ_y	μ	σ_ϵ	δ	ρ_L	ρ_H	$\mathbb{V}$
0.52	43.33	63.99	9.84	25.53	13.07	0.27	0.45	0.72	499.72
(0.08)	(2.49)	(1.67)	(0.80)	(5.81)	(4.32)	(0.06)	(0.09)	(0.07)	(31.95)

NOTE: White standard errors are below estimates. Standard errors for the equilibrium objects are computed using the delta method. β is fixed at 0.85.

Second, our model allows for a disciplining effects of elections if $\delta > 0$ as an equilibrium outcome. Suppose instead that none of the bad governors exerted high effort so that $\delta = 0$. This would be a case of only adverse selection with no moral hazard, since types are linked to effort deterministically. If this were the case, then the difference between the performance of a reelected governor across terms would display two properties: It would have a zero mean and it would be a normally distributed variable with no skewness.[27] In the data, $y_2 - y_1$ has a mean of -1.98, though it is not significantly different from zero, but it is highly negatively skewed with a skewness of -1.77.

Given the discussion so far, we can summarize how the parameters in our model are identified as follows. Suppose that we have a set of governors where we know their performance and reelection outcomes. Furthermore, let us assume some values for each of the structural parameters, which also pin down the equilibrium value of δ. With these in hand, we can assign probabilities to each governor of being a given type and exerting a given level of effort. Then σ_y will adjust to match the dispersion of performance conditional on effort level. Next, performance conditional on high and low effort pins down Y_H and Y_L. Next π and implicitly δ will adjust to match $corr(y_1, y_2)$ and distribution of $y_2 - y_1$. Finally, the election shock distribution will adjust to match the variation in election outcomes given performance.

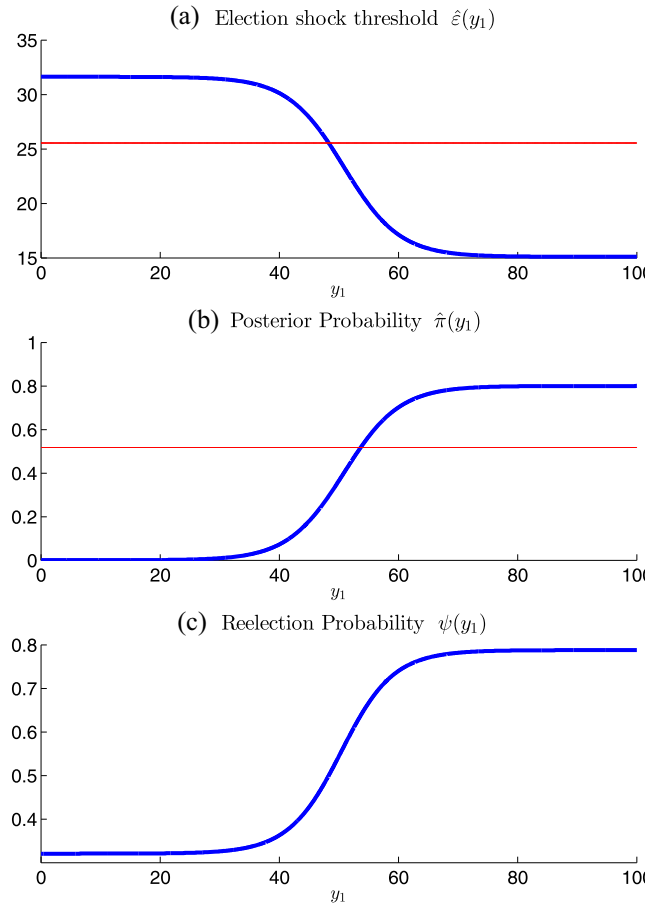
5. ESTIMATION RESULTS

5.1. *Benchmark Model.*

5.1.1. *Basic results.* The estimates of the six structural parameters and the four equilibrium outcomes are given in Table 4. Several things can be noted. A total of 52% of governors in our sample are good, and, based on the standard error, we strongly reject the two extremes, all governors being good or all governors being bad. Of the bad governors, 27% of them exert high effort in their first term and thus are disciplined. This is also highly statistically significant. Exerting high effort (for any governor) leads to an average increase in performance ($Y_H - Y_L$) of over 20 JAR points, which is highly significant, both statistically and economically. High effort increases the probability of reelection from 45% to 72%. The mean of the election shock is 25.5, and it is highly significant. The election shock threshold $\hat{\varepsilon}(y_1)$, the posterior probability that an incumbent's type is good $\hat{\pi}(y_1)$, and the reelection probability $\psi(y_1)$, all conditional on observed y_1, are illustrated in Figure 1. The shapes of all these mappings originate from the shape of the $\hat{\pi}(y_1)$ mapping, which in turn uses the normality of the process that determines y_1. A small first-term JAR, for example, 25, signals to the voter that the governor did not exert high effort; as a result he assigns a near-zero probability of the governor being the good type. Then, for this governor to win reelection she needs an election shock of around 32 or larger. Since

This correlation is significantly negative at -0.59, and it is largely driven by large declines in y_2 relative to y_1 by bad governors.

[27] To see this, first note that for a bad governor both y_1 and y_2 are drawn from $N(Y_L, \sigma_y^2)$, and for a good governor they are drawn from $N(Y_H, \sigma_y^2)$. Given that they are i.i.d., the difference $(y_2 - y_1) \sim N(0, 2\sigma_y^2)$ in both cases, and given the properties of the normal distribution it would have no skewness.



NOTES: The horizontal reference lines in the first and second panels are μ (the mean of the election shock process) and π (the unconditional probability of a governor being good), respectively.

FIGURE 1

EQUILIBRIUM MAPPINGS [COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

this is quite reasonable given the estimated values of $\mu = 25.5$ and $\sigma_\varepsilon = 13.1$, there is about a 30% probability for this governor to win reelection, despite poor first-term performance. As y_1 increases so does $\hat{\pi}(y_1)$, until y_1 hits 80, after which the reelection probability remains constant at around 80%, reflecting the possibility of an unfavorable election shock after a very strong performance in the first term.

5.1.2. *Model fit.* To understand the implications of our estimated model and how it fits the data, we compute some results and discuss them in this section. Although some of these can be computed analytically, many cannot, and thus we resort to simulations, where we simulate the model for 1,000,000 hypothetical governors.[28]

First, turning to the moments we report in Table 3, our model displays a good fit. To highlight a few key ones, $corr(y_1, y_2|R = 1) = 0.34$ in our model, which, as we explained in Section 4.4, is key for our finding that there are two types of governors. Similarly $y_2 - y_1$ is negative on

[28] Here and throughout the article lifetime welfare is computed as $[1/(1-\beta)]\mathbb{E}(y)$ since, given linear utility, this is equivalent to the more obvious definition of welfare $\mathbb{E}[\sum_{t=0}^{T} \beta^t y_t]$. Here the $\mathbb{E}(.)$ operator is over all possible random events including the type of governors. Note that our definition of welfare excludes ε, and since the definition of $\mathbb{V}$ in (7) includes ε in the utility of the voters, these numbers are not identical.

average, with a negative skewness (albeit smaller than the data), which indicates evidence that there are disciplined bad governors.

Our estimation results indicate the presence of large election shocks with a positive mean—the 95% probability range for election shocks is −0.1 to 51.1. To put this in perspective, the 95% confidence intervals for performance conditional on high and low effort are 24.0 to 62.6 and 44.7 to 83.3, respectively. Note further that the average of the election shock μ, which is 25.5, exceeds the average effect of high versus low effort, that is $Y_H - Y_L$, which is 20.7. The large value for σ_ε is explained by the presence of many governors in our sample who get reelected despite low performance or who lose reelection despite high performance. Turning to μ, though one may be tempted to think of $\mu > 0$ as reflecting *incumbency advantage*, this is not necessarily the case.[29] Even if $\mu = 0$, there could be an electoral advantage or disadvantage associated with incumbency: Election outcomes are determined as in (12), and $\mathbb{V}$ depends in complicated ways on all structural parameters and the behavioral response of governors, that is, whether or not bad governors exert high effort. Thus, instead of focusing on the sign or magnitude of μ, a useful discussion here is to demonstrate the degree of incumbency advantage in our data, as μ will adjust in conjunction with other parameters to match this. There are two ways of measuring the incumbency advantage in our data. First, 61.3% of incumbents (57 out of 93) in our data win reelection. Second, reelection surprises in our data favor incumbents: If we look at governors who have $y_1 < Y_L$ but win reelection and those with $y_1 > Y_H$ but who lose reelection, the former is 30.0%, whereas the latter is 15.4%. As Table 3 shows, our model matches these numbers closely, considering the small number of governors used in computing the data moments for the last two rows of Table 3.

Table A-1 in the Online Appendix lists the individual governors in our sample, their performance, and some potentially interesting statistics that can be computed from our estimation. In particular, we show the performance measures y_1 and y_2 that go into the estimation, $\hat{\pi}(y_1)$, the updated probability that the governor is a good type after observing y_1, $\psi(y_1)$, the probability that the governor will win reelection given her first-term performance, as well as a new measure $\bar{\pi}(y_1, R, y_2)$, which shows the ex post assessment of a governor's type that one could calculate after having observed her performance in both terms and the reelection outcome. In Online Appendix E we show the exact expression for $\bar{\pi}(y_1, R, y_2)$, and we provide three examples to illustrate how our model works by using these measures.

We can also talk about how good a fit our model provides to the data. A useful way to do this is by looking at $\psi(y_{1k})$, the model's implied probability that an incumbent k will win reelection after observing y_{1k}, which is reported in Table A-1. If we select a rule that predicts reelection whenever $\psi(y_{1k}) > 0.5$, then we can correctly predict the reelection outcomes for 75% of the governors (49 wins and 21 losses) in our sample, incorrectly predicting only 15 wins and 8 losses. One way to assess the performance of a probability forecast such as $\psi(y_1)$ is to use the Brier (1950) score, which is defined as $(1/n)\sum_{k=1}^{n}[\psi(y_{1k}) - R_k]^2$, where $R_k \in \{0, 1\}$ is the election outcome. The Brier score is between 0 (a perfect prediction) and 1, with smaller numbers indicating a better forecast. Our model gets a Brier score of 0.195. For comparison, a naive forecast that uses the overall fraction of governors who win in our sample (61.3%) for *each* governor instead of the $\psi(y_{1k})$ measure gets a Brier score of 0.237.[30] A Diebold and Mariano (1995) test, as in Lahiri and Yang (2013), rejects equal accuracy between our model's forecast and the naive forecast with a p-value of 0.01. Our model places quite a bit of structure on the relationship between the observable variables (reelection outcomes and first-term JAR in this case), which in principle puts it at a disadvantage against a reduced-form model like a probit. However, an estimated probit model that uses JAR in the first term as a predictor (i.e., a reduced-form model using the same observables as our structural model) yields *lower*

[29] We define incumbency advantage as the unconditional probability of winning reelection for an incumbent, $\mathbb{P}(R = 1)$, being greater than 0.5, and explore the dependence of reelection probability on various factors in Online Appendix D where we show how $\mathbb{P}(R = 1)$ changes under different parameter configurations.

[30] As a point of comparison, a naive forecast that the incumbent wins 50% of the time would lead to a Brier score of 0.25.

TABLE 5
SOME PROPERTIES OF THE ESTIMATED MODEL AND VARIOUS COUNTERFACTUALS

	Benchmark	One-Term	No-Discipline	All-Discipline
Good governors in term 1	51.8%	51.8%	51.8%	51.8%
Good governors in term 2	59.6%	–	63.0%	51.8%
Good governors overall	54.8%	51.8%	56.3%	51.8%
High effort in term 1	64.7%	51.8%	51.8%	100.0%
High effort in term 2	59.6%	–	63.0%	51.8%
High effort overall	62.8%	51.8%	56.3%	81.3%
Average performance in term 1 (JAR points)	56.7	54.0	54.0	64.0
Average performance in term 2 (JAR points)	55.6	–	56.3	54.0
Average performance overall (JAR points)	56.3	54.0	55.0	60.1
Welfare	375.3	360.3	366.5	400.8
(% of benchmark)	–	−4.0	−2.3	6.8

NOTE: The numbers in this table are obtained by simulating the model under various assumptions for 1,000,000 hypothetical governors, given the structural parameters in Table 4, except where noted. The third and fourth columns are obtained by setting the cost of exerting high effort for bad governors to $c = 1$ and $c = 0$, respectively, and re-solving the model so that the voters optimally react to this.

prediction accuracy, that is, a higher Brier score (though the difference is no longer statistically significant).

Finally, we also provide evidence that our model and the estimated parameters are sensible by considering a small set of governors that served under a one-term limit regime, for whom we have JAR data. The one-term version of our model is very simple: Good governors exert effort, bad governors do not. We expect the average performance to be normally distributed with mean $\pi Y_H + (1-\pi)Y_L$ and standard deviation $\sqrt{\pi(1-\pi)(Y_H - Y_L)^2 + \sigma_y^2}$, or 54.02 and 14.26, respectively. Getting reliable JAR data on one-term-limited governors is challenging since most states who had such limits had them before our JAR data starts, with the exception of Virginia, which is the sole state that has this limit. Using a small data set of 30 governors with JAR data from 11 states the average JAR is 55.30 and the standard deviation is 12.11. These numbers are remarkably close to the implied numbers from our model, especially given the small sample, and empirical distribution tests fail to reject the hypothesis that the data come from the particular normal distribution implied by our model. Thus we conclude that, at the very least, the one-term-limited governor data do not refute our model and the estimation.[31]

5.2. *Measuring the Effects of Elections.* Elections have three consequences in our model: discipline (bad governors exert high effort to secure reelection), selection (more good than bad governors are reelected), and mimicking (bad governors who are disciplined look like good governors). To measure the first two effects, we compare the outcomes in the benchmark model with a counterfactual model where governors can only serve one term. Having more disciplined governors improves first-term outcomes relative to the one-term case since more governors overall will exert high effort. In turn, when there is a second term, the selection effect can be measured as the improvement in outcomes in the second term of the benchmark model relative to the one-term counterfactual model, as good governors have a higher reelection rate than bad governors and then they exert high effort in their second term.

These effects are not independent of each other. To see this, consider the case where all bad governors are disciplined (all-discipline counterfactual in Table 5), which we implement by setting $c = 0$ for all bad governors. In this version, all governors, good or bad, exert high effort in their first term. As a result, performance is no longer an informative signal for screening governors, thus leading to identical fractions of each type of governor across the first and second

[31] An obvious question to ask is if we can estimate the one-term-limited model. Unfortunately, the answer is no because the data provide two pieces of information (the mean and variance of the JAR distribution) but there are four parameters to pin down: π, Y_H, Y_L, and σ_y.

terms, so that the percentages in each term would be identical to the one-term counterfactual. This means the outcome in the second term will be identical to the one-term outcome, that is, there is no selection effect. It is important to realize that the lack of selection is a negative consequence of having more disciplined governors in the first term. We call this third effect "mimicking" and, to remember that it makes outcomes worse, use a negative sign. Thus, we distinguish between "pure selection," which is the screening effect of elections were there no mimicking, and "selection," as defined above. Naturally, selection is equal to pure selection plus the (negative) effect of mimicking.

To identify pure selection, we consider a second counterfactual, one where there is no discipline as an equilibrium outcome (no-discipline counterfactual in Table 5). To obtain this, we assume that the cost of exerting high effort for bad governors is $c = 1$, which means none of them exerts high effort. This ensures that $\delta = 0$ in equilibrium and Equation (4) no longer is a part of the description of equilibrium. Now, when we compare the second-term performance in the no-discipline version with the one-term case, this yields a measure of pure selection since all governors play their types and voters are better able to screen bad governors. Furthermore, the difference between pure selection and selection (which is equal to the difference between second-term performance in the no-discipline case and the benchmark) is a measure of mimicking.

To conduct these counterfactuals—the one-term and no-discipline counterfactuals—we re-solve our model under the appropriate assumptions, using the structural parameters we estimated. Naturally, in each case the voter solves his problem taking these new assumptions into account, and this influences all equilibrium mappings, including, for example, the reelection rule, where relevant, and thus the equilibrium outcomes ρ_L, ρ_H, and $\mathbb{V}$. The simulation results for these counterfactual are presented in Table 5.

Table 5 shows the fraction of good governors, fraction of governors that exert high effort, and average performance in the two terms and also in the aggregate. These numbers are not very meaningful by themselves, and below we compare them to the results from various counterfactuals to identify the discipline and selection effects precisely. The table also shows the lifetime expected welfare of the electorate in the estimated model, which is 375.3. To put this in perspective, the "first-best" in this economy, one where in every term high effort is exerted, leads to a lifetime expected welfare of $(1 - \beta)^{-1} Y_H = 426.6$. This shows that our benchmark economy has about 12% less welfare than this ideal one. Even the all-discipline counterfactual, which has welfare 6.8% better than the benchmark, achieves a level of welfare that is about 6% lower than the first-best because of low performance in the second term and randomness of elections.

Table 6 shows two different approaches to computing the three effects in question: discipline, selection, and mimicking. The first approach, labeled A, uses the change in the fraction of high-effort governors, measured in percentage points, whereas the second approach, labeled B, uses the change in performance, both as absolute change in performance and also as relative to the counterfactual as we explain now. Comparing the benchmark version with the one-term case, we find a 12.9 percentage point increase in the fraction of governors exerting high effort in their first term, which leads to an increase of 2.7 JAR points in performance, or a 4.9% increase. These are our measures of discipline. The selection effect is lower in magnitude, namely, a 7.8 percentage point increase in the fraction of high-effort governors in the second term relative to the one-term case, leading to a 1.6 JAR point or 2.9% increase in performance. However, the improvement in the second term due to selection is partially cancelled due to mimicking: a 3.4 percentage point decline in the fraction of high-effort governors in the second term, leading to a 0.7 JAR point or 1.2% decline in performance. We use bootstrapping methods to compute confidence intervals, and all the estimates in panel (a) of Table 6 are significant at the 5% level.[32]

[32] The changes using JAR points seem smaller than those that use change in the fraction of governors exerting high effort. This is because of the stochastic relationship between effort and performance.

TABLE 6
DISCIPLINE AND SELECTION

(a) Measures of Discipline and Selection	
Discipline A : Change in fraction of high-effort governors in term 1 (benchmark versus one-term)	12.9**
Discipline B : Change in performance in term 1 (benchmark versus one-term)	2.7 [4.9%]**
Selection A : Change in fraction of high-effort governors in term 2 (benchmark versus one-term)	7.8**
Selection B : Change in performance in term 2 (benchmark versus one-term)	1.6 [2.9%]**
Mimicking A : Change in fraction of high-effort governors in term 2 (benchmark versus no-discipline)	−3.4**
Mimicking B : Change in performance in term 2 (benchmark versus no-discipline)	−0.7 [−1.2%]**

(b) Comparison	Value
Discipline A − Selection A	4.9**
Discipline B − Selection B	1.0 [1.9]**
Discipline A − Pure Selection A	1.4
Discipline B − Pure Selection B	0.3 [0.7]

NOTE: The numbers in this table are obtained by simulating the model for 1,000,000 hypothetical governors, given the structural parameters in Table 4. The $\delta = 0$ version is solved assuming $c = 1$. In panel (a), all changes in fractions (such as the ones for Discipline A, Selection A, and Mimicking A measures) are reported as percentage point changes. In panel (b), Pure Selection is defined as Selection minus Mimicking, where Mimicking is negative to emphasize its welfare-reducing nature. Numbers in square brackets are in percentage point units and show the difference between the corresponding term in square brackets in panel (a). Numbers in panel (b) may not exactly correspond to the differences in the numbers in panel (a) due to rounding. In both panels ** denotes significance at the 5% level.

In panel (b), we compare the measures of discipline and selection. When we allow mimicking to reduce the gains due to selection, both types of measures, A and B, show that the effect of discipline is significantly different from selection—an almost 5 percentage point difference in the fraction of good governors in term 1 versus term 2 that leads to about a 2 percentage points increase in performance in term 1 versus in term 2. However, when we compare discipline with pure selection, the differences are not significant at the 10% level.

Returning to the discussion of the reduced-form estimates for discipline and selection in Section 2, results in this section demonstrate the advantages of using a structural approach. In Section 2, a simple regression of JAR-based performance on an election-eligible dummy would lead one to conclude that there was no significant effect on governor performance of being election-eligible. We argued that this coefficient will be the combination of the three effects we measured in this section and that it was not possible to identify all three individually. Table 6 shows that all three effects are practically and statistically significant.

5.3. *Other Counterfactuals.* In addition to the two counterfactuals we used in the previous section (the no-discipline and all-discipline counterfactuals), we consider two counterfactual exercises where we change various aspects of the environment. In each case we solve the model with these changes, holding the structural parameters at their estimated values. The counterfactuals capture changes in term limits to a single term and the possibility of a signal of governor effort.

5.3.1. *Term limits.* The first counterfactual we consider is a change in the term limits to a single term. In principle, the model could be extended to any term limit n (though it becomes more complicated and tedious).[33] There are two key problems with extending the analysis to

[33] Doing so takes some additional work—for example, now governors will take in to account the outcome y_{i-1} in making their effort decision e_i in term i, or more generally the voter's perception of the governor's type in term i, since this influences election outcomes in addition to the performance y_i.

values of n larger than two, both of which have to do with things we do not explicitly model. First, we do not model the choice of entering into politics and do not have a model-based explanation of why certain fractions of governors are good or bad. We think that changing term limits may influence self-selection into politics and thus may alter π. Second, we use the election shock ε to capture everything that may affect election outcomes other than the performance of the incumbent governor, and this shock is a crucial input to the reelection probability for the incumbent governor. The parameters that govern it, μ and σ_ε, are estimated using our sample of governors with a two-term limit. It is reasonable to expect that with a different term limit these parameters will change. As it turns out there are no U.S. states that currently have a term limit greater than two, and thus we cannot use any information to reliably discipline values for μ and σ_ε for each term of the incumbent. (Utah briefly had a three-term limit in the 1990s; no states are expected to enact term limits of three terms of higher.) Similarly, $n = 1$ and $n = 2$ are the most commonly used limits in the United States historically and presently (aside from having no term limits at all), and this limits the appeal of considering higher values for n. In contrast, for the $n = 1$ case, we only need π, Y_L, Y_H, and σ_y, and using the estimated parameters from the benchmark case will not create a major problem. In fact, our results in Section 5.1.2 show that our model with $n = 1$ is very much consistent with the limited data we have where governors have a one-term limit.

Results for the model with a one-term limit are reported on the second column of Table 5. Since there is no reelection, only good governors would exert high effort, leading to an average performance of 54 JAR points. Lifetime voter utility is 360.3 in this case. Comparing these results with the benchmark model, having a two-term instead of a one-term limit is unambiguously better for the voter. First of all, more governors exert effort in their first terms, leading to a higher average JAR. This is because 26.8% of the bad governors exert effort in addition to all the good governors, leading to high effort 64.7% of the time in the first term, compared with 51.8% in the one-term case. This increases average JAR in the first term from 54 to 56.7. Second, because a higher fraction of bad than good governors are screened out in elections, more governors are good in the second term: 59.6% relative to the unconditional probability of 51.8%. Since these governors always exert effort, the average JAR in the second term is 55.6, compared to 54 in the one-term case. Putting these together, the lifetime voter utility goes up from 360.3 to 375.3, which is a 4.2% increase. Put differently, a voter in a two-term regime would be willing to give up about 2.3 JAR points *every* term ad infinitum in order to remain in that regime and not switch to a regime of one-term limits. It is also important to note that the voter is better off in the two-term regime because the governors' performance in both terms is higher relative to the case of a one-term limit. Using the no-discipline counterfactual also shown in Table 5 we can decompose this welfare difference between the one-term and the two-term regimes. About two-thirds of it is due to the disciplining effect of elections: Going from the one-term-limit counterfactual to the no-discipline counterfactual, welfare goes up from 360.3 to 366.5 (all of which is due to selection), whereas from the no-discipline counterfactual to the benchmark model welfare goes up from 366.5 to 375.3.

5.3.2. *Noisy effort signal.* In our second counterfactual, we consider providing a (noisy) signal to the electorate about the effort of the governors. This will help in understanding the importance of the election shock for the strength of discipline effects as well as the trade-off between discipline and selection. The extension of the model was presented in Section 3.6. Table 7 reports discipline and selection measures (analogous to Table 6) for different values of the partially and fully informative signals of governor effort, the latter in both the presence and absence of an election shock. Throughout this section, we assume the structural parameters shown in Table 4 are unchanged but solve for the equilibrium objects for every probability ζ considered. We show the recomputed δ in the table.

The first column shows the benchmark results of no effort signal discussed above, which correspond to $\zeta = 0.5$ in this version. The second column shows the effect of a partially informative signal of effort, $\zeta = 0.75$. Relative to the case of an uninformative (or no) signal, the fraction

TABLE 7
RESULTS FROM THE VERSION WITH AN EFFORT SIGNAL

	$\zeta = 0.5$	$\zeta = 0.75$	$\zeta = 0.9$	$\zeta = 1$	$\zeta = 1$ and $\sigma_\epsilon = 0$
δ	0.27	0.30	0.35	0.42	1.00
Discipline B	4.9%	5.5%	6.5%	7.8%	18.4%
Selection B	2.9%	2.5%	1.6%	0.0%	0%
Welfare gain	–	0.1%	0.4%	0.5%	4.8%

NOTE: The first column ($\zeta = 0.5$) shows the benchmark results from Tables 4 and 6. Structural parameters are kept as in Table 4. See Table 6 for the definitions of the discipline and selection measures. The "B" measures are reported as percentage change from the value in a world with no reelection. Welfare gain is relative to the benchmark with no signal (or equivalently with $\zeta = 0.5$).

of bad governors disciplined rises from 27% to 30%. This is consistent with what theory would lead us to expect: A higher probability of observing "shirking" leads to more bad types exerting high effort. Selection is smaller at 2.5% instead of 2.9%. Better information makes incumbent performance a more informative signal. This should improve both discipline and pure selection. Whether selection improves or declines, however, depends on how the increased mimicking compares to the increased pure selection as we discussed in Section 5.2. Our estimates suggest that mimicking increases faster than pure selection, leading to a decrease in selection.

The next column shows the effects of an increase in ζ to 0.9, which means the signal is correct 90% of the time. Now 35% of bad governors are disciplined, and selection falls further relative to the no-signal case.

To better understand the magnitudes of these effects, we also considered the case of $\zeta = 1$, that is, perfect observability of effort, as shown in the fourth column of Table 7. (This, of course, is not equivalent to perfect observability of type, since bad governors still can, and do, mimic the effort levels of good governors.) We see that the fraction of bad governors disciplined in their first term rises to 42%, an increase by more than half of the 27% when effort was unobservable, but not by more as one might be inclined to expect. The reason why full observability of effort does *not* lead to all bad types exerting high effort in their first term is the existence of the election shock. Even if a governor is known to be of bad type—perfectly indicated in this case by low effort—she can still win reelection with a sufficiently positive realization of ε (her reelection probability is $\rho_L = 0.23$); conversely, even if a bad type exerts high effort, she is not guaranteed reelection (her reelection probability is $\rho_H = 0.66$) if the realization of ε is sufficiently negative. Therefore, bad types with a sufficiently high draw of c will still find it optimal to exert low effort, even though it will be fully apparent to the voter that they did so. Hence, discipline is mitigated by the randomness of reelection outcomes due to reasons unrelated to performance, as theory once again would suggest. Turning to selection, what we called "pure selection" is virtually fully cancelled by mimicking, and there is a negligible selection effect (zero up to rounding error).

To confirm our conjecture that the lack of full discipline is due to the presence of the election shock, we solve the model with full observability of effort ($\zeta = 1$) and with $\sigma_\varepsilon \approx 0$, so that the election shock is constrained to take its mean value $\mu = 25.5$. There is no longer the possibility of a very positive realization of ε to "save" a low-effort incumbent. Now all bad governors exert high effort, and all are reelected. Mimicking of good governors by all bad governors implies that there is *no* selection effect, and the fraction of good governors in the second term is identical to the fraction in the first term since all incumbents win reelection.

The last row in Table 7 shows how the welfare of the voter changes in each case. Having a moderately informative effort signal is worth an extra 0.1% of welfare to the voter whereas making effort fully observable leads to an improvement of 0.5%. As discipline is increasing, selection falls, and on net the increase in welfare is small. These gains pale in comparison to the one in the last column where in addition to making effort fully observable, we eliminate the uncertainty of elections by removing the election shock. This achieves a welfare improvement of 4.8% over the two-term benchmark. This gain is smaller than the 6.8% we report for the

TABLE 8
ROBUSTNESS OF ESTIMATION RESULTS

	Benchmark	All Surveys	No Election Year	Year-by-Year Average
π	0.52	0.50	0.53	0.53
	(0.08)	(0.10)	(0.09)	(0.09)
δ	0.27	0.26	0.26	0.29
	(0.06)	(0.06)	(0.07)	(0.07)
Discipline B	4.9%	4.8%	4.6%	5.3%
Selection B	2.9%	2.8%	2.7%	3.2%
Welfare gain	4.2%	4.0%	3.9%	4.5%
		Median JAR	Minimum JAR	Keep Undecided
π		0.54	0.45	0.46
		(0.09)	(0.12)	(0.12)
δ		0.25	0.23	0.23
		(0.07)	(0.06)	(0.06)
Discipline B		4.5%	6.4%	4.3%
Selection B		2.9%	3.5%	2.3%
Welfare gain		3.9%	5.3%	3.5%
		Adjusted JAR	Unemployment Rate	Income Growth
π		0.52	0.42	0.59
		(0.09)	(0.09)	(0.19)
δ		0.26	0.23	0.09
		(0.06)	(0.08)	(0.05)
Discipline B		4.6%	1.9%	1.7%
Selection B		2.8%	0.9%	1.5%
Welfare gain		3.9%	1.5%	1.6%

NOTE: The top of each panel shows the reestimated π and δ for each case with standard errors in parentheses. See Table 6 for the definitions of the discipline and selection measures. Reported welfare gains are relative to the one-term regime.

all-discipline counterfactual in Section 5.2. This is because without election shocks, all governors serve two terms and bad governors are able to play their type. With election shocks, some incumbents are replaced by first-term governors who always exert high effort. This means in this case there are fewer second -terms on average, which improves welfare.

Our results show that greater transparency would not *in itself* significantly increase effort, especially due to the randomness of election outcomes (that is, their dependence on other factors). If greater transparency also made election outcomes themselves less stochastic, they could increase effort (through more discipline) and thus welfare significantly, as suggested by our exercise in the last column of Table 7.

5.4. *Robustness.* In this section, we perform three types of exercises to explore the robustness of our parameter estimation. First, as explained in Section 4.3, we made some choices in preparing the JAR data for estimation. Here we consider some alternative choices. Second, we use the "adjusted JAR," which strips the JAR data from what may be considered as partisan effects, as we also explain in Section 4.3. Third, we use two state-wide macroeconomic indicators as performance measures instead of JAR. The results are reported in Table 8, where in the interest of space, we only report our new estimates for π and δ as well as one measure for discipline, selection, and welfare.

Our benchmark measure of governor performance averaged the results of all JAR surveys over a governor's first term up to and including June of the election year, where we used the fraction of respondents who classify the governor as excellent or good out of those who express an opinion (that is, eliminating the undecided respondents). The first six columns in Table 8 following the Benchmark report the results from six alternative choices: using all surveys in the

first term up to the election (All Surveys), dropping all surveys taken in the election year (No Election Year), taking the average JAR in each year of the term and then taking the year-by-year average so that respondent sentiment in a year with many surveys would not be overweighted (Year-by-Year Average), using the median (Median JAR) or the minimum JAR (Minimum JAR) instead of the average, and taking the fraction of respondents who classified the governor as excellent or good out of *all* respondents including the undecided (Keep Undecideds), which essentially classifies the undecided as expressing low approval. As the estimates make clear, the results are robust to all of these alternative performance calculations. The key is that the identification of π and δ is not affected by these variations—roughly half of all governors are bad, and of those about a quarter of them are disciplined.

The next column shows the estimation results with the "adjusted JAR." Results are virtually identical to the benchmark results. This is not surprising because the adjustments tend to be small—the median adjustment is about 2 JAR points. The last two columns show the estimation results with macroeconomic variables instead of JAR as measures of governor performance. First, we should note that the state unemployment rate is by far the most important determinant of JAR (that we could identify), and it is a reasonable predictor of reelection performance as we explained above. State per capita personal income growth has a smaller correlation with both JAR and reelection outcomes. Both versions show smaller discipline and selection effects relative to our benchmark (about a third), but discipline is still larger than selection. The estimated value of δ in the version with unemployment rate is similar to the benchmark value at 0.23, but in the version with income growth δ is small and statistically indistinguishable from zero at 5% significance. When we inspect the detailed results when using state-level macroeconomic indicators, we see, however, that stochastic outcomes play a much larger role. (See also footnote 15.) To see this, consider the ratio $(Y_H - Y_L)/\sigma_y$ as a measure of the importance of luck—the numerator is the expected performance differential from exerting high effort, and the denominator is one standard deviation of the election shock. This ratio is about 2.1 in our benchmark model, and it is 2.2 in the version with the unemployment rate. This means exerting high effort is expected to create about twice the increase in performance relative to one standard deviation increase in luck. In the income growth version, this ratio is only 0.4, indicating that luck is much more important. We find this an unattractive feature of using these narrower measures of performance, further strengthening the view presented in Section 4.3 of the preferability of using a multifaceted measure of performance such as JAR.

6. CONCLUSIONS

In this article, we constructed a political agency model with adverse selection and moral hazard, and we structurally estimated the model. The aim was to disentangle the various effects that electoral accountability has on policymaker performance—specifically discipline and selection effects—and, more generally, to assess the empirical relevance of the widely used political agency model.

Our structural approach provides an alternative to reduced-form approaches that seek to separately identify and estimate discipline and selection effects. We estimate the effects on the performance of U.S. governors under the common two-term limit regime relative to the counterfactual case where reelection is not allowed, so that elections can neither discipline nor allow selection based on performance. Our structural approach also allows counterfactual experiments to assess the welfare effects of electoral accountability under different configurations of governor incentives and voter information.

We find a significant discipline effect of reelection incentives as well as a somewhat weaker selection effect. Quantifying these effects allows us to assess their relative importance. More generally, our results indicate that a formal political agency model stressing the role of accountability finds support in the data, an important point given the widespread use of the political agency approach in theoretical political economy models. We further find that electoral accountability has important welfare implications, where the possibility of reelection induces a

significant increase in welfare due to its inducing higher effort. We should note, as we made clear in the article, that we consider only the discipline and selection effects of elections and the implied utility effects. There may be other effects—loss of experience induced by term limits versus allowing "new blood" to be injected into politics—that may have significant utility implications. However, a tractable structural model requires focusing on a limited number of issues, and we believe that those we chose are first order.

Further research may help address interesting questions raised by these results. Why is there such a large fraction of "bad" governors in the data? Why do not reelection incentives discipline a larger fraction of them? Arguing that there is a large stochastic element to elections does not fully answer the second question. These two questions are likely related. For example, it is reasonable to think about the link between the importance of rents in office and self-selection into politics.

In our opinion, structural estimation can be quite helpful in gaining a deeper understanding of issues of politician performance and electoral accountability. We believe this article is a useful step in that direction.

Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure A.1: Game Tree
Table A.1: Governors
Table A.2: Incumbency Advantage and Structural Parameters

REFERENCES

Alt, J., E. Bueno de Mesquita, and S. Rose, "Disentangling Accountability and Competence in Elections: Evidence from U.S. Term Limits," *Journal of Politics* 73 (2011), 171–86.

Ashworth, S., "Electoral Accountability: Recent Theoretical and Empirical Work," *Annual Review of Political Science* 15 (2012), 183–201.

Avis, E., C. Ferraz, and F. Finan, "Do Government Audits Reduce Corruption? Estimating the Impacts of Exposing Corrupt Politicians," *Journal of Political Economy* 126 (2018), 1912–64.

Banks, J., and R. Sundaram, "Adverse Selection and Moral Hazard in a Repeated Elections Model," in W. Barnett, M. J. Hinich, and N. J. Schofield, eds., *Political Economy: Institutions, Information, Competition and Representation* (New York, NY: Cambridge University Press, 1993) 295–311.

Barro, R., "The Control of Politicians: An Economic Model," *Public Choice* 14 (1973), 19–42.

Bernhardt, M. D., and D. E. Ingberman, "Candidate Reputations and the 'Incumbency Effect,'" *Journal of Public Economics* 27 (1985), 47–67.

Besley, T., *Principled Agents? The Political Economy of Good Government* (Oxford: Oxford University Press, 2007).

———, and A. Case, "Does Electoral Accountability Affect Economic Policy Choices? Evidence from Gubernatorial Term Limits," *Quarterly Journal of Economics* 110 (1995), 769–98.

———, and ———, "Political Institutions and Policy Choices: Evidence from the United States," *Journal of Economic Literature* 41 (2003), 7–73.

Beyle, T., R. Niemi, and L. Sigelman, "Gubernatorial, Senatorial, and State-level Presidential Job Approval Ratings: A Compilation of Data," *State Politics and Policy Quarterly* 2 (2002), 215–29.

Brender, A., and A. Drazen, "Political Budget Cycles in New versus Established Democracies," *Journal of Monetary Economics* 52 (2005), 1271–95.

———, and ———, "How Do Budget Deficits and Economic Growth Affect Reelection Prospects? Evidence from a Large Panel of Countries," *American Economic Review* 98 (2008), 2203–20.

Brier, G. W., "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review* 78 (1950), 1–3.

Dal Bó, E., and M. Rossi, "Term Length and the Effort of Politicians," *Review of Economic Studies* 78 (2011), 1237–63.

DeBacker, J., "The Price of Pork: The Seniority Trap in the US House," *Journal of Public Economics* 95 (2011), 63–78.

DIEBOLD, F. X., AND R. MARIANO, "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics* 13 (1995), 253–63.
DIERMEIER, D., H. ERASLAN, AND A. MERLO, "A Structural Model of Government Formation," *Econometrica* 71 (2003), 27–70.
DRAZEN, A., AND E. Y. OZBAY, "Does 'Being Chosen to Lead' Induce Non-Selfish Behavior? Experimental Evidence on Reciprocity," *Journal of Public Economics* 174 (2019), 13–21.
DUGGAN, J., AND C. MARTINELLI, "Electoral Accountability and Responsive Democracy," mimeo, 2015.
FEARON, J., "Electoral Accountability and Control of Politicians: Selecting Good Types versus Sanctioning Poor Performance," in A. Przeworski, B. Stokes and S. C. Manin, eds., *Democracy, Accountability, and Representation* (Cambridge: Cambridge University Press, 1999) 55–97.
FEREJOHN, J., "Incumbent Performance and Electoral Control," *Public Choice* 50 (1986), 5–26.
FERRAZ, C., AND F. FINAN, "Electoral Accountability and Corruption in Local Governments: Evidence from Audit Reports," *American Economic Review* 101 (2011), 1274–311.
FINAN, F., AND M. MAZZOCCO, "Electoral Incentives and the Allocation of Public Funds," NBER Working Paper No. 21859, 2016.
GAGLIARDUCCI, S., AND T. NANNICINI, "Do Better Paid Politicians Perform Better? Disentangling Incentives from Selection," *Journal of the European Economic Association* 11 (2013), 369–98.
GOWRISANKARAN, G., M. F. MITCHELL, AND A. MORO, "Electoral Design and Voter Welfare from the US Senate: Evidence from a Dynamic Selection Model," *Review of Economic Dynamics* 11 (2008), 1–17.
HEALY, A. J., N. MALHOTRA, AND C. H. MO, "Irrelevant Events Affect Voters' Evaluations of Government Performance," *Proceedings of the National Academy of Sciences* 107 (2010), 12804–9.
JACOBSON, G. C., "'The Polls': Polarized Opinion in the States: Partisan Differences in Approval Ratings of Governors, Senators, and George W. Bush," *Presidential Studies Quarterly* 36 (2006), 732–57.
LAHIRI, K., AND L. YANG, "Forecasting Binary Outcomes," in A. Timmermann and G. Elliott, eds., *Handbook of Economic Forecasting*, Volume 2B (Amsterdam: North Holland, 2013) 1025–106.
LIST, J., AND M. STURM, "How Elections Matter: Theory and Evidence from Environmental Policy," *Quarterly Journal of Economics* 121 (2006), 1249–81.
LUCAS, R., "Econometric Policy Evaluation: A Critique," in K. Brunne, and A. Meltzer, eds., *The Phillips Curve and Labor Markets* (Amsterdam: North Holland, 1976), 19–46.
MASKIN, E., AND J. TIROLE, "The Politician and the Judge: Accountability in Government," *American Economic Review* 94 (2004), 1034–54.
MERLO, A., "Bargaining over Governments in a Stochastic Environment," *Journal of Political Economy* 105 (1997), 101–31.
SIEG, H., AND C. YOON, "Estimating Dynamic Games of Electoral Competition to Evaluate Term Limits in U.S. Gubernatorial Elections," *American Economic Review* 107 (2017), 1824–57.
SMART, M., AND D. M. STURM, "Term Limits and Electoral Accountability," *Journal of Public Economics* 107 (2013), 93–102.
STRÖMBERG, D., "How the Electoral College Influences Campaigns and Policy: The Probability of Being Florida," *American Economic Review* 98 (2008), 769-807.
VLAICU, R., AND A. WHALLEY, "Hierarchical Accountability in Government," *Journal of Public Economics*, 134 (2016), 85–99.
WOLFERS, J., "Are Voters Rational? Evidence from Gubernatorial Elections," mimeo, Stanford University, 2007.