Inter-American Development Bank

A Structural Model of Electoral Accountability

Department of Research and
Chief Economist

S. Borağan Aruoba
Allan Drazen
Razvan Vlaicu

DISCUSSION
PAPER Nº
IDB-DP-515

May 2017

A Structural Model of Electoral Accountability

S. Borağan Aruoba*
Allan Drazen**
Razvan Vlaicu***

* University of Maryland
** University of Maryland, National Bureau of Economic Research and Centre for Economic Policy Research
*** Inter-American Development Bank

May 2017

Inter-American Development Bank

Abstract[*]

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

JEL classifications: D72, D73, H70, C57

Keywords: Discipline, Selection, Political agency, Elections, structural estimation, Maximum likelihood

[*] Aruoba (Corresponding author, aruoba@econ.umd.edu): University of Maryland. Drazen (drazen@econ.umd.edu): University of Maryland, NBER, and CEPR. Vlaicu (vlaicu@iadb.org): Inter-American Development Bank. The authors thank Jim Alt, Tim Besley and Shanna Rose for their assistance with data and general feedback, Ethan Kaplan, Nuno Limão, 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-Jiménez for research assistance at various stages of the project.

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], Ferejohn [1986]). 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]).[1]

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. In order to identify the effects of elections they rely on variation in electoral incentives induced by term limits, either across electoral terms, e.g., comparing reelection-eligible to lame-duck officials, or across electoral regimes, e.g., comparing officials serving under shorter and longer term limit regimes.

In this paper 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, which displays both adverse selection and moral hazard, to measure the importance of discipline and selection quantitatively. We also perform several counterfactual exercises to assess the welfare implications of term limits and 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

[1] 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).

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 the effort of governors), 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 good and exert 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 2/3 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 rather than 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.[2]

The plan of the paper is as follows. In the next section we discuss the 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

[2] 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.

in Section 5. The final section presents conclusions. An 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.[3] For example, Besley and Case (1995, 2003), Besley (2006) and Alt, Bueno de Mesquita and Rose (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 which may reflect both discipline and selection, as these authors also recognize.

Some of the above research makes further assumptions to try to disentangle the effects. For example, Besley (2006) 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 rather than selection.

A related approach is proposed by Alt, Bueno de Mesquita and Rose (2011) who ar-

[3] 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.

gue 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 can not (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 $baseline + discipline$ while 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 using the data we use

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 OLS. The unit of observation is 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 while 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.

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 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 term. The coefficient γ captures the 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 to 6), then we get a positive coefficient which 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 all first-termers become lame ducks, thus the coefficient measures a mixture of discipline and mimicking. The structural approach we propose below will be able to separate these two effects.

2.3 Structural Estimation

There are very few papers that use a structural rather than reduced form approach to study the effects of elections on policymaker performance and policy outcomes.[4] Sieg and Yoon (2016) ask whether electoral competition leads U.S. governors to moderate their fiscal policies: Democratic incumbents to act more fiscally conservative, Republican incumbents to act more fiscally liberal. They find this is the case for about 1/5 of Democratic incumbents and 1/3 of Republican incumbents. Our paper differs in two key respects. We study the effects of reelection on governor effort and overall performance in office rather than on their fiscal stance. The papers are thus complementary in focusing on different outcome aspects. Second, our model considers both the moral hazard and adverse selection problems of the electoral agency, whereas their model focuses on adverse selection abstracting from governors' effort decision.

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 percent of funds are misallocated relative to the social optimum, and study the welfare effects of alternative electoral institutions such as approval voting. While the paper is also concerned with the effect of electoral accountability on outcomes, the mechanisms it highlights – interaction among legislators and their decisions

[4] Structural estimation is relatively rare in political economy. Some examples are Merlo (1997), Diermeier, Eraslan, and Merlo (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's actions are probabilistic and not strategic as in our model.

to run for office – are quite different than ours, as is the outcome studied (efficient versus inefficient allocation of public spending). They also look at effects on legislators (and their interactions) rather than chief executives in states.

Finally, Avis, Ferraz, and Finan (2016) 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 non-electoral 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 non-electoral explaining 94 percent 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 non-electoral rather than electoral.

Given the differences in outcomes considered (as in Sieg and Yoon, 2016) or in mechanisms for accountability and levels of government (as in Finan and Mazzocco, 2016 and Avis, Ferraz, and Finan, 2016), we view our paper 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. (In section 5.1.2 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 paper.) 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, while bad governors have a positive utility cost c, which is expressed as a fraction of the rents r of office.[5] For ease of exposition, we define $c(e;\theta)\, r$ the cost of effort level e for a governor of type θ, where

$$c(H;G) = c(L;G) = c(L;B) = 0 \text{ and } c(H;B) = c \tag{2}$$

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 in office.[6] 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 (as specified in equations (6) 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

$$\max_{e_1,e_2} [1 - c(e_1;\theta)]\, r + [\mathbf{1}_H \rho_H + (1 - \mathbf{1}_H)\, \rho_L]\, [1 - c(e_2;\theta)]\, r \tag{3}$$

where e_i is effort in term i and $\mathbf{1}_H$ is an indicator which 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 ($e_1 = H$)

[5] 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 towards voters, as discussed in Drazen and Ozbay (2015). 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.

[6] 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.

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.[7]

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.[8] To derive e_1, note that if 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 forego 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

$$c < \rho_H - \rho_L \tag{4}$$

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

$$\delta = \mathbb{P}(c < \rho_H - \rho_L) = \rho_H - \rho_L \tag{5}$$

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

$$y_i|\,(e_i = H) \quad \sim \quad N\left(Y_H, \sigma_y^2\right) \tag{6a}$$

$$y_i|\,(e_i = L) \quad \sim \quad N\left(Y_L, \sigma_y^2\right) \tag{6b}$$

for term $i = 1, 2$, where $Y_H > Y_L$. 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. Note that we also assume that the relationship between effort and performance is independent of the governor's type or the

[7] If we assumed that good types like exerting high effort, i.e. $c(H;G) < 0$, 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.

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

term she is in.

We further assume probabilistic voting in that the utility of the voter is affected by a shock $\varepsilon \sim N\left(\mu, \sigma_{\varepsilon}^{2}\right)$ 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 capture 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 5.1.2.

Define $W\left(y_1, \varepsilon\right)$ 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

$$W\left(y_1, \varepsilon\right) = y_1 + \beta \max_{R \in \{0,1\}} \mathbb{E}\left\{R\left[y_2 + \varepsilon + \beta W\left(y_1', \varepsilon'\right)\right] + (1-R)\, W\left(y_1', \varepsilon'\right) | y_1, \varepsilon\right\} \tag{7}$$

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 $W\left(y_1', \varepsilon'\right)$ 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 $\left(y_1, \varepsilon\right)$ 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 can rewrite the voter's problem as

$$W\left(y_1, \varepsilon\right) = y_1 + \beta \max_{R \in \{0,1\}} \left\{R\left[\mathbb{E}\left(y_2 | y_1\right) + \varepsilon + \beta \mathbb{V}\right] + (1-R)\, \mathbb{V}\right\} \tag{8}$$

where we use $\mathbb{V}$ to denote $\mathbb{E}\left[W\left(y_1', \varepsilon'\right)\right]$, 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

$$\begin{aligned}\mathbb{V} &= [\pi + (1-\pi)\,\delta]\,\frac{1}{\sigma_y\sigma_\varepsilon}\int\int W\left(y_1', \varepsilon'\right)\phi\left(\frac{y_1' - Y_H}{\sigma_y}\right)\phi\left(\frac{\varepsilon' - \mu}{\sigma_\varepsilon}\right)dy_1' d\varepsilon' \\ &\quad + (1-\pi)\,(1-\delta)\,\frac{1}{\sigma_y\sigma_\varepsilon}\int\int W\left(y_1', \varepsilon'\right)\phi\left(\frac{y_1' - Y_L}{\sigma_y}\right)\phi\left(\frac{\varepsilon' - \mu}{\sigma_\varepsilon}\right)dy_1' d\varepsilon'\end{aligned} \tag{9}$$

where $\phi(\cdot)$ represents the standard normal PDF. Equation (9) makes it explicit that there is 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

$$\mathbb{E}\left(y_2|y_1\right) = \hat{\pi}\left(y_1\right)Y_H + \left[1 - \hat{\pi}\left(y_1\right)\right]Y_L \tag{10}$$

where $\hat{\pi}\left(y_1\right) \equiv \mathbb{P}\left(\theta = G|y_1\right)$, that is, the voter's posterior probability that the incumbent is good after observing first-term performance. Using (10) we can write $W\left(y_1, \varepsilon\right)$ as

$$W\left(y_1, \varepsilon\right) = y_1 + \beta \max_{R\in\{0,1\}}\left[R\left\{\hat{\pi}\left(y_1\right)Y_H + \left[1-\hat{\pi}\left(y_1\right)\right]Y_L + \varepsilon + \beta\mathbb{V}\right\} + (1-R)\,\mathbb{V}\right] \tag{11}$$

3.4 Election

If types were observable, the voter would reelect only good governors since they would exert high effort in their second term while bad governors would not. Since voters only observe performance y_1, not type or effort, their reelection rule is related to performance. However, the relationship is probabilistic, not deterministic, because the election shock might change the voter's performance-based assessment of the incumbent. Solving the discrete choice problem in (11), the incumbent would win reelection, i.e. $R = 1$, if and only if

$$\hat{\pi}\left(y_1\right) > \frac{(1-\beta)\,\mathbb{V} - Y_L - \varepsilon}{Y_H - Y_L} \tag{12}$$

which 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. We can summarize the voting rule

$R(y_1, \varepsilon)$ with the following

$$R(y_1, \varepsilon) = \begin{cases} 0 & \text{if } \varepsilon \leq \hat{\varepsilon}(y_1) \\ 1 & \text{if } \varepsilon > \hat{\varepsilon}(y_1) \end{cases} \tag{13}$$

where $\varepsilon = \hat{\varepsilon}(y_1)$ characterizes the points (y_1, ε) for which (12) holds with equality with

$$\hat{\varepsilon}(y_1) = (1-\beta)\mathbb{V} - \hat{\pi}(y_1)(Y_H - Y_L) - Y_L \tag{14}$$

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

$$\begin{aligned} \hat{\pi}(y_1) &\equiv \mathbb{P}(\theta = G|y_1) = \frac{\mathbb{P}(\theta = G)\, p(y_1|\theta = G)}{p(y_1)} \\ &= \frac{\pi\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)}{[\pi + (1-\pi)\delta]\,\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) + (1-\pi)(1-\delta)\,\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right)} \end{aligned} \tag{15}$$

where δ, as defined in (5), 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.

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

$$\begin{aligned} \psi(y_1) &\equiv \mathbb{P}(R = 1|y_1) = \mathbb{P}[\varepsilon > \hat{\varepsilon}(y_1)] \\ &= 1 - \Phi\left[\frac{\hat{\varepsilon}(y_1) - \mu}{\sigma_\varepsilon}\right] \end{aligned} \tag{16}$$

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

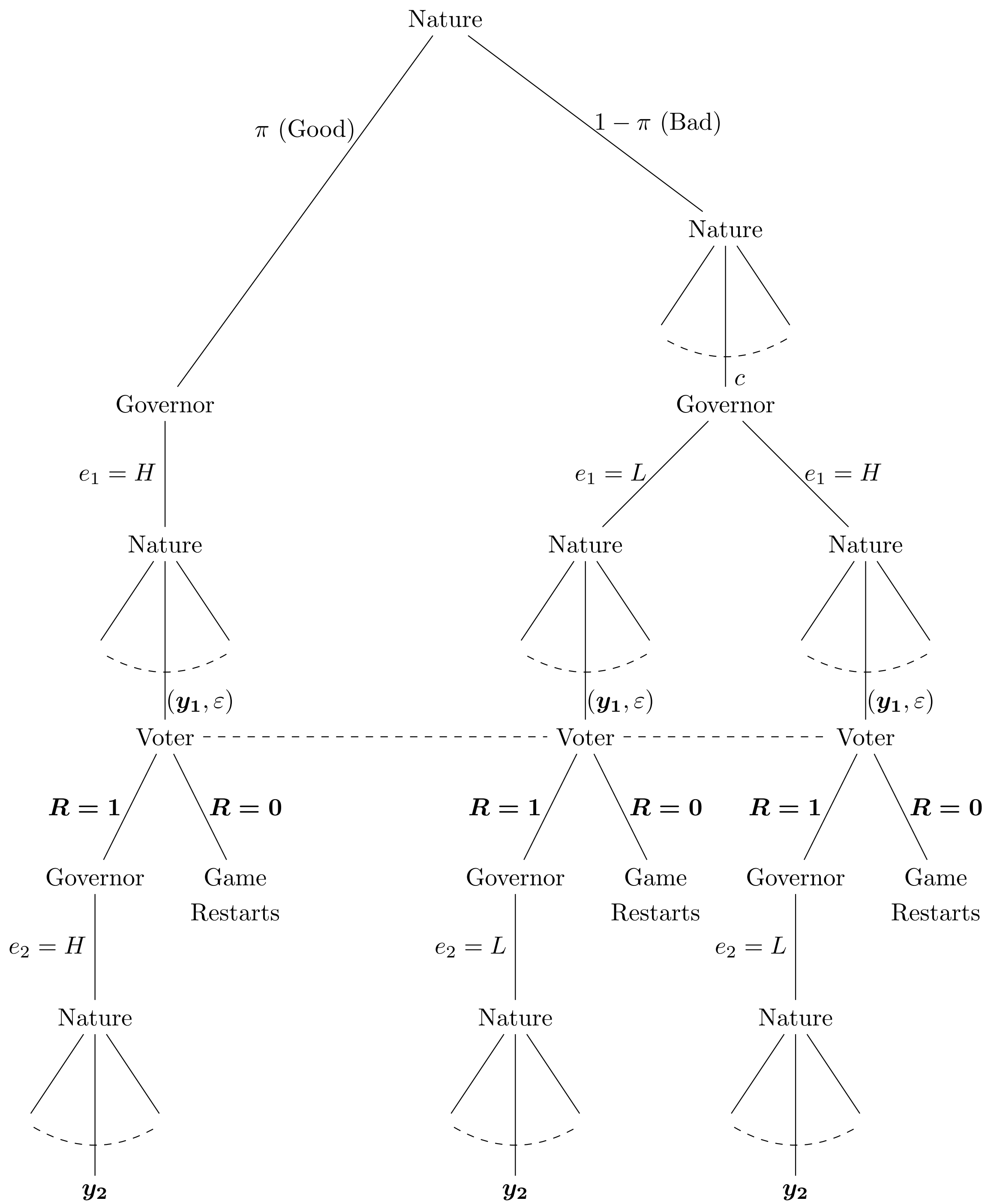
Finally, the last piece we need is the probabilities ρ_L and ρ_H that the governor was taking as given. These can be obtained by integrating $\psi(y_1)$ with respect to the performance distributions as in

$$\rho_H = \frac{1}{\sigma_y}\int \psi(y_1)\,\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) dy_1 \tag{17}$$

$$\rho_L = \frac{1}{\sigma_y}\int \psi(y_1)\,\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right) dy_1 \tag{18}$$

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

Figure 1: **Game Tree**



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 shock ε is realized.

4. 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.

5. 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.

6. At the end of the term, a new election is held between two randomly-drawn candidates and the game restarts.

3.5 Equilibrium

A strategy for a governor is a choice of whether or not to exert high effort, i.e. $e_i(c) \in \{H, L\}$, 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, i.e. $R(y_1, \varepsilon) \in \{0, 1\}$, 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 *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 Appendix. ■

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 (3) 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 (11) 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 life-time 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 (4) holds; the voter reelects the incumbent if and only if (12) holds; and voter beliefs about the incumbent's type are given by (15).*

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

$$\zeta \equiv \mathbb{P}\left\{z_1 = H | e_1 = H\right\} = \mathbb{P}\left\{z_1 = L | e_1 = L\right\} \tag{19}$$

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

$$\begin{aligned}
\hat{\pi}(y_1, z_1) \quad &\equiv \quad \mathbb{P}\left(\theta = G | y_1, z_1\right) = \frac{\pi p\left(y_1, z_1 | \theta = G\right)}{\pi p\left(y_1, z_1 | \theta = G\right) + (1-\pi)\, p\left(y_1, z_1 | \theta = B\right)} \\
&= \begin{cases} \dfrac{\pi \zeta \phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)}{\left[\pi + (1-\pi)\,\delta\right] \zeta \phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) + (1-\pi)(1-\delta)(1-\zeta)\,\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right)} & \text{if } z_1 = H \\ \dfrac{\pi (1-\zeta)\,\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)}{\left[\pi + (1-\pi)\,\delta\right](1-\zeta)\,\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) + (1-\pi)(1-\delta)\,\zeta \phi\left(\frac{y_1 - Y_L}{\sigma_y}\right)} & \text{if } z_1 = L \end{cases}
\end{aligned} \tag{20}$$

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

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

$$\rho_H = \frac{1}{\sigma_y} \int \left[\zeta \psi\left(y_1, H\right) + (1-\zeta)\,\psi\left(y_1, L\right)\right] \phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) dy_1 \tag{21}$$

$$\rho_L = \frac{1}{\sigma_y} \int \left[(1-\zeta)\,\psi\left(y_1, H\right) + \zeta \psi\left(y_1, L\right)\right] \phi\left(\frac{y_1 - Y_L}{\sigma_y}\right) dy_1 \tag{22}$$

Further details are presented in the Appendix.

4 Solution, Estimation, and Data

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

4.1 Solution

The model has seven structural parameters: $\pi, \beta, Y_H, Y_L, \sigma_y, \mu$, 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 $\psi(y_1)$. We solve for the equilibrium as follows.

The first thing to notice is that once $\mathbb{V}$ and δ are known, ρ_H and ρ_L follow from (17) and (18), with $W(y_1, \varepsilon)$, $\hat{\pi}(y_1)$, $R(y_1, \varepsilon)$, $\hat{\varepsilon}(y_1)$, and $\psi(y_1)$ obtained using (11), (15), (13), (14), and (16), respectively. Thus solving for the equilibrium amounts to satisfying (5) and (9). Define two residuals $\mathcal{R}_1$ and $\mathcal{R}_2$ as the differences between conjectures for δ and $\mathbb{V}$ and the model-implied values from (5) and (9), respectively

$$
\begin{aligned}
\mathcal{R}_1 &\equiv \delta - \rho_H + \rho_L && (23)\\
\mathcal{R}_2 &\equiv \mathbb{V} - [\pi + (1-\pi)\delta] \frac{1}{\sigma_y \sigma_\varepsilon} \int\int W(y_1', \varepsilon')\, \phi\left(\frac{y_1' - Y_H}{\sigma_y}\right) \phi\left(\frac{\varepsilon' - \mu}{\sigma_\varepsilon}\right) dy_1' d\varepsilon' && (24)\\
&\quad - (1-\pi)(1-\delta) \frac{1}{\sigma_y \sigma_\varepsilon} \int\int W(y_1', \varepsilon')\, \phi\left(\frac{y_1' - Y_L}{\sigma_y}\right) \phi\left(\frac{\varepsilon' - \mu}{\sigma_\varepsilon}\right) dy_1' d\varepsilon'
\end{aligned}
$$

where equilibrium requires $\mathcal{R}_1 = \mathcal{R}_2 = 0$. This yields 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. To recap, then, given the seven structural parameters, we find the equilibrium values for $\mathbb{V}$ and δ. With these nine values in hand, we can compute everything we need, including the mappings $\hat{\pi}(y_1)$, $\hat{\varepsilon}(y_1)$, $R(y_1, \varepsilon)$, $W(y_1, \varepsilon)$ and $\psi(y_1)$ that we will use for computing the likelihood function, to which we turn next.

4.2 Estimation

We estimate the structural parameters using Maximum Likelihood. Our data set will consist 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. For a governor who wins reelection, we observe the triplet $(y_1, R = 1, y_2)$. For a governor who loses reelection, we observe the pair $(y_1, R = 0)$. Each of these outcomes might come from different combinations of governor types, effort choices and reelection shocks. The density of a generic governor winning reelection while producing performance of y_1 and y_2 can be obtained as

$$\begin{aligned} p_W(y_1, y_2) \equiv\ & \frac{1}{\sigma_y^2}\left[\pi\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)\psi(y_1)\phi\left(\frac{y_2 - Y_H}{\sigma_y}\right)\right. \\ & +(1-\pi)\delta\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)\psi(y_1)\phi\left(\frac{y_2 - Y_L}{\sigma_y}\right) \\ & \left.+(1-\pi)(1-\delta)\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right)\psi(y_1)\phi\left(\frac{y_2 - Y_L}{\sigma_y}\right)\right] \end{aligned} \tag{25}$$

The three terms capture the cases where the governor is good, bad but disciplined, and bad and not disciplined, respectively. Similarly, the density of a governor of unspecified type losing reelection with first-term performance of y_1 is given by

$$\begin{aligned} p_L(y_1) \equiv\ & \frac{1}{\sigma_y}\left[\pi\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)[1-\psi(y_1)]\right. \\ & +(1-\pi)\delta\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)[1-\psi(y_1)] \\ & \left.+(1-\pi)(1-\delta)\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right)[1-\psi(y_1)]\right] \end{aligned} \tag{26}$$

For a governor k with (y_{1k}, R_k, y_{2k}), we compute her contribution to the log-likelihood using

$$\log L_k = R_k \log\left[p_W(y_{1k}, y_{2k})\right] + (1 - R_k)\log\left[p_L(y_{1k})\right] \tag{27}$$

and the log-likelihood is simply given by

$$\log \mathcal{L} = \sum_{k=1}^{n} \log L_k \tag{28}$$

Estimating the structural parameters requires maximizing $\log \mathcal{L}$, which we do using standard numerical optimization routines. We estimate six structural parameters $(\pi, Y_L, Y_H, \sigma_y, \mu, \sigma_\varepsilon)$ and fix $\beta = 0.85$, which represents roughly a 4% annual discount rate over a four-year term.[9] Once estimates for the structural parameters are obtained, estimates for equilibrium outcomes $(\rho_H, \rho_L, \delta, \mathbb{V})$ 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

In order 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. (2016) 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 section 5.1.3. We therefore want a broader measure of performance that might possibly capture the multi-faceted nature of performance.[10] Nevertheless, for completeness sake, in Section 5.3, as a part of our robustness checks, we show our estimation results using 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

[9] We tried alternative values of β and the estimates of structural parameters do not change and only the equilibrium object $\mathbb{V}$ adjusts.

[10] 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.

scarce. We chose to use job approval ratings (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, Niemi, and Sigelman (2002), and we update their dataset through the end of 2014 using various online resources. Potential voters are asked to rate the governor as "excellent", "good", "fair" and "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.[11] 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 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.[12] The second set are variables measuring partisanship in the state: state's population (possibly capturing homogeneity of preferences in smaller states), whether the governor is of the same party as the U.S. president, "partisan fit" which shows match between the party of the governor and how the state voted in the previous presidential election and the percent of voters of the governor's or the opposing party.[13] The third set of variables are governor characteristics: age, gender, years of education, whether the governor is a lawyer, or served in the military. The regressions also include state and year fixed effects.

The first column in Table 2 uses only measures of state economic performance. The

[11] 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.

[12] 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.

[13] 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.

Table 2: **Determinants of Job Approval Ratings**

	(1)	(2)	(3)	(4)
	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 regression include state and year fixed effects. It uses 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 shows 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.

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 come 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, while 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 they perform 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 the results of the 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 (8 years) in age between two otherwise identical governors creates 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 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.[14] Some of these narratives are presented in Section 5.1.3. 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.[15]

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 we do not capture more than half of what is in JAR with this regression. Given the sensitivity checks we have done, we are confident that it is the multi-faceted 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 non-performance-related factors.

Finally, we explain how we convert individual JAR survey results to measures that cover terms of governors. In order 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, i.e., 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

[14] 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, Malhotra, and Mo [2010]), but this is consistent with our modeling of the election shock ε.

[15] 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 rather than reward fiscal manipulation in election years.

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 (2006), including the outcomes of their reelection bids.[16] 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.

- Drop governors who did not have any term limits, or had a one-term limit or a three-term limit.[17]
- Drop governors during whose stints state election laws regarding term limits changed.
- 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.
- 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.[18] 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 limits early on in our sample, except for one governor from the 1960s, our data covers 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%

[16] 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 A1).

[17] In principle using data from states with a one-term limit as either a part of the estimation or as an "out-of-sample" verification could be a good idea. As it turns out, one-term limits were most common before 1980 (except for Virginia which still continues the practice) and JAR data for these governors are not available. Hence, sparsity of data prevents us from using these cases.

[18] 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.

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.[19] We provide the basic data that we use for estimation, namely (y_1, R, y_2) in Table A1 in the Appendix.

5 Estimation Results

5.1 Benchmark Model

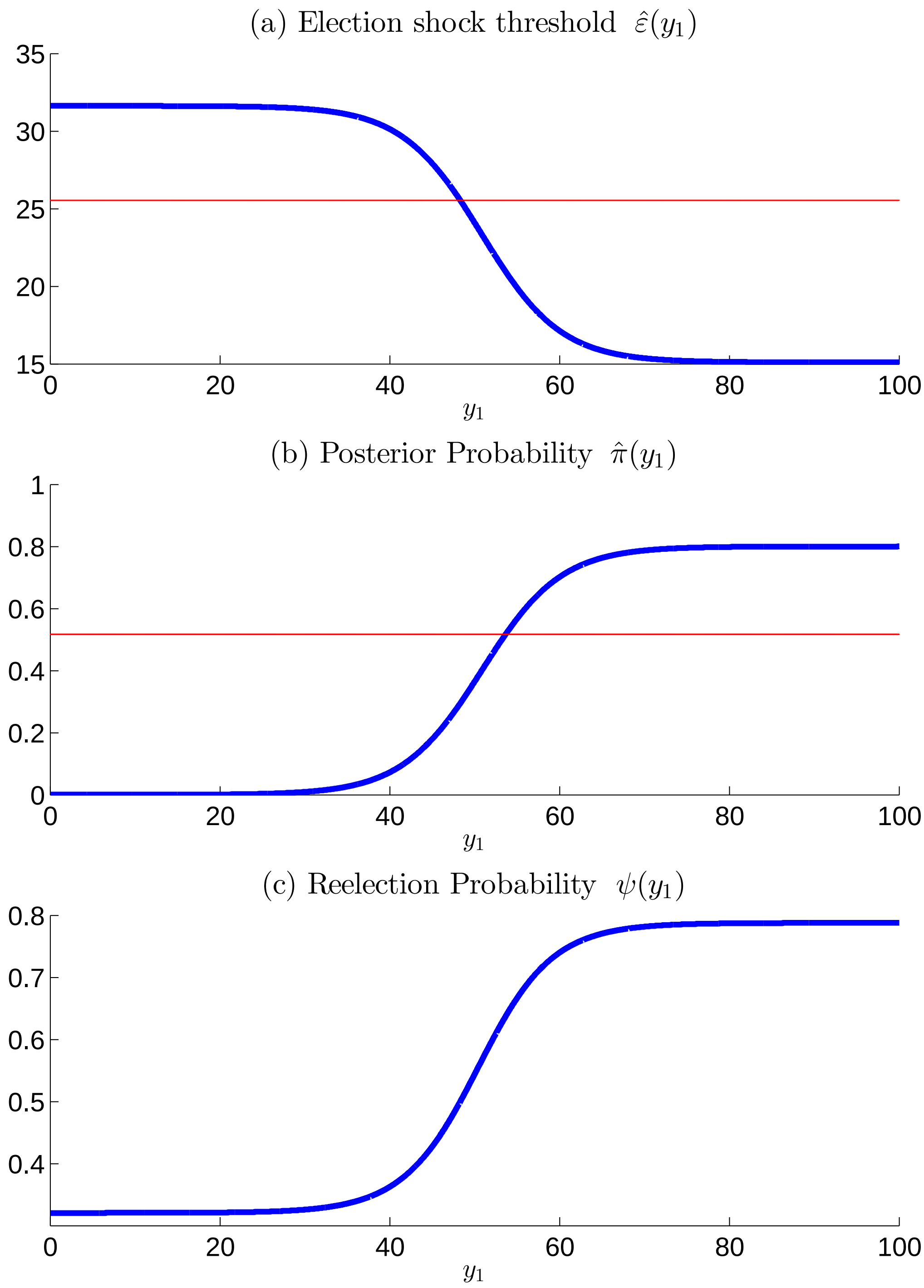
5.1.1 Basic Results

The estimates of the six structural parameters and the four equilibrium outcomes are given in Table 3. Several things can be noted. 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 2. 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 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.

In the specification of our model we assumed the variance of outcomes was independent

[19] Our largest sample of 588 governor stints, which includes governors we dropped with the filters above show 96% 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.

Figure 2: **Equilibrium Mappings**



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.

Table 3: **Parameter Estimates**

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

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.

of effort. When we allow the $y|e$ distribution to have a variance that depends on e, this unrestricted model has a log-likelihod which is only 0.23 log-points higher and there are no substantive changes in the results. We conclude that the restriction we place by assuming a single variance is not rejected by the data, and given the small improvement in goodness of fit, a model selection criterion that favors parsimony (e.g. Schwarz Information Criterion) chooses our restricted model.

5.1.2 Identification

Before we turn to the implications of our model with the particular parameter estimates, it will be useful to discuss their identification. There are three key features of the parameter estimates we discuss in turn. First, we find very strong evidence that there are two types of governors $(0 < \pi < 1)$. Second, we also find very strong evidence that some fraction of bad governors are disciplined $(\delta > 0)$. Third, we find the presence of large election shocks. We demonstrate how our model identifies these three features by considering counterfactuals.

First, consider the possibility that there was only one type of governor. If $\pi = 0$ then there are only bad governors, while 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.[20] In our data, however, this correlation is 0.36 with a standard deviation of 0.14 (significant at the 5% level) and our model delivers 0.39 over the long simulation explained in Section 5.1.3, which is very close.

Next, suppose 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.[21] In the data $y_2 - y_1$ has a mean of -1.98, though it is not significantly different from zero, and the Jarque-Bera test of normality is overwhelmingly rejected, primarily because of negative skewness. Our model delivers a mean of -2.68 and a negative skewness.

Finally, our estimation results indicate the presence of large election shocks with a positive mean – the 95% confidence interval 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. We explain how σ_ε and μ are identified in turn. If σ_ε were equal to 0, but we kept the estimated value for μ, then any governor with $\hat{\varepsilon}(y_1) < \mu$ would have to win reelection, and all others would have to lose – reelection would become a deterministic function of y_1. However in our sample there are many governors who lose reelection despite good first-term performance, as well as those who win despite bad performance, implying the presence of election shocks. More specifically, if we used a rule of predicting reelection when $\hat{\varepsilon}(y_1) < \mu$, out of the 93 governor stints in our sample, we would correctly predict the reelection outcome of 70 of them. We would, however, incorrectly predict reelection for 15 of them and incorrectly predict losing reelection for 8 of them. This shows that there need to be large election shocks beyond what is simply implied by μ.

Turning to μ, we note first that though one may be tempted to think of *incumbency*

[20] It is important to note that the cases where $\pi = 0$ and $\pi = 1$ are not covered by our model since updating of types as in (15) 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.

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

advantage as being characterized by $\mu > 0$, this is not necessarily the case. Even if $\mu = 0$ there could be an electoral advantage or disadvantage associated with incumbency: election outcomes are determined as in (15) and $\mathbb{V}$ depends in complicated ways on all structural parameters and the behavioral response of governors, i.e. whether or not bad governors exert high effort. 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 Appendix D where we show how $\mathbb{P}(R = 1)$ changes under different parameter configurations. Thus, rather than focusing on the identification 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, 61.2% 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 $y_1 > Y_H$ but lose reelection and those with $y_1 < Y_L$ but win reelection, the former is 15% while the latter is 30%.

Once these key parameters or equilibrium objects are pinned down, the other three structural parameters, that is, Y_H, Y_L and σ_y, follow from matching other properties of the JAR data. These include the mean and variance of y_1 and y_2.

5.1.3 Measures of interest

To understand what the parameter estimates imply for our model with a two-term limit, we report some results in Table 4. While some of the measures can be computed analytically, many cannot, and thus we resort to simulations where we simulate the model for 1,000,000 hypothetical governors.[22]

Table 4 shows the fraction of good governors, fraction of governors that exert high effort and the average performance in the two terms and also in the aggregate. These numbers are not very meaningful by themselves and in the next section we compare them to the results from various counterfactuals to identify the discipline and selection effects precisely. The table also shows the life-time 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 every

[22]Here and throughout the paper life-time welfare is computed as $[1/(1-\beta)]\,\mathbb{E}(y)$ since given linear utility this is equivalent to the obvious definition of welfare $\mathbb{E}\left[\sum_{t=0}^{T}\beta^t y_t^s\right]$. Here the $\mathbb{E}(.)$ operator is over all possible random events including the type of governors. In practice we compute $\mathbb{E}(y)$ via a simulation with 1,000,000 governors. Note that our definition of welfare excludes ε and since the definition of $\mathbb{V}$ in (9) includes ε in the utility of the voters, these numbers are not identical.

Table 4: **Some Properties of the Estimated Model**

Good governors in Term 1	51.8%
Good governors in Term 2	59.6%
Good governors Overall	54.8%
High effort in Term 1	64.7%
High effort in Term 2	59.6%
High effort Overall	62.8%
Average Performance in Term 1 (JAR Points)	56.7
Average Performance in Term 2 (JAR Points)	55.6
Average Performance Overall (JAR Points)	56.3
Life-time Discounted Welfare for Voter	375.3

Note: The numbers on this table are obtained by simulating the model for 1,000,000 governors, given the structural parameters in Table 3.

term high effort is exerted, leads to a life-time 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.

Table A1 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, as well as $\hat{\pi}(y_1)$, 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 statistic

$$\bar{\pi}(y_1, R, y_2) \equiv \mathbb{P}(\theta = G | y_1, R, y_2) = \begin{cases} \dfrac{\pi \frac{1}{\sigma_y^2} \phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) \psi(y_1) \phi\left(\frac{y_2 - Y_H}{\sigma_y}\right)}{p_W(y_1, y_2)} & \text{if } R = 1 \\ \dfrac{\pi \frac{1}{\sigma_y} \phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) [1 - \psi(y_1)]}{p_L(y_1)} & \text{if } R = 0 \end{cases} \tag{29}$$

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.

We present three examples to illustrate how our model works in terms of disciplining effects and election shocks. The first is Guy Hunt, who was the governor of Alabama between 1987 and 1993. His first-term performance is 60.1, which is slightly lower than Y_H but sufficiently far away from Y_L for voters to believe that he was a good type with 70% probability just prior to his reelection bid. This implies a 74% chance that he would win the election. He won a second term, but had performance of only 38.2 in the second term. As a result his

$\bar{\pi}(y_1, R, y_2)$ is only 8%. According to our model, he was therefore probably a bad governor who exerted high effort in the first term and low effort in his second term. Incidentally, he was forced to resign towards the end of his second term because he was convicted of theft, conspiracy and ethics violations, including taking money from his first-term inaugural fund for personal use. Our second example is Lawton Chiles, who was the governor of Florida between 1991 and 1999. His first-term JAR was only 38.7 leading voters to believe according to the model that he was a good type with only 35% probability. This corresponds to the high disapproval rate Floridians gave him after he cut funds for education in his first budget and especially after what was seen as his inadequate response to Hurricane Andrew. Nonetheless he narrowly won reelection against his challenger Jeb Bush in 1994 after an extremely heated campaign and a massive get-out-the-vote telephone effort the day before the election (which, it was later admitted, spread false information about Bush).[23] This would be an example of a low-performing first-term governor winning reelection due to what the model characterizes as a positive election shock. Our third example is Roy Barnes, who was elected governor of Georgia in 1998, but lost reelection in 2002. His first-term performance was 77.8 which would correspond to voters thinking he was a good type with probability of 80%. The model gives him a reelection probability of 79% consistent with what pre-election polls said: his challenger, Sonny Perdue, never came closer than seven percentage points behind him, with Barnes having a lead of eleven points in the final poll taken before the election. His loss was considered an enormous upset. It has been attributed to unexpectedly high turnout of voters angry about his positions on education reform and his overhaul of the state flag, issues made salient by his opponent in the campaign. (Hayes and McKee, 2004). This would be an example of a large negative realization of the election shock.

Finally, we can also talk about how good a fit our model provides to the data. In Table A1 we report $\psi(y_{1k})$, the model's implied probability that an incumbent k will win reelection after observing y_{1k}. 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 zero (a perfect prediction) and one, 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 (61.2%) for *each* governor instead of

[23] http://www.nytimes.com/1994/10/19/us/1994-campaign-florida-governor-jeb-bush-debate-exchanging-taunts-claims.html

the $\psi(y_{1k})$ measure gets a Brier score of 0.237.[24] A Diebold-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* prediction accuracy, that is, a higher Brier score (though the difference is no longer statistically significant). We conclude that our structural model explains more of the variation in governor performance than the non-structural alternatives considered.

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). In order 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, which means that 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 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

[24] 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, regardless of the outcome.

of elections were there no mimicking, and selection as defined above. Naturally, selection is equal to pure selection plus the (negative) effect of mimicking.

In order to identify pure selection, we consider a second counterfactual, one where there is no discipline as an equilibrium outcome. 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 (5) no longer is a part of the description of equilibrium.

In order to conduct these counterfactuals – the one-term and no-discipline counterfactuals – we re-solve our model with 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 the one-term counterfactual are presented in Section 5.4.1, while those for the no-discipline counterfactual are presented in Section 5.4.2.

Table 5 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, while 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 effect of selection is lower in magnitude, namely a 7.8 percentage point increase in the fraction of high-effort governors in the second term, 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 – 3.4 percentage point decline in 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 5 are significant at the 5% level.[25]

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 – almost 5 percentage point difference in fraction of good governors in term 1 versus term 2 which leads to about a 2

[25] 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 5: **Discipline and Selection**

(a) Measures of Discipline and Selection

Discipline A : Change in Fraction of High-Effort Governors in Term 1 (Benchmark vs. 1-Term)	12.9**
Discipline B : Change in Performance In Term 1 (Benchmark vs. 1-Term)	2.7 [4.9%]**
Selection A : Change in Fraction of High-Effort Governors in Term 2 (Benchmark vs. 1-Term)	7.8**
Selection B : Change in Performance In Term 2 (Benchmark vs. 1-Term)	1.6 [2.9%]**
Mimicking A : Change in Fraction of High-Effort Governors in Term 2 (Benchmark vs. $\delta = 0$)	-3.4**
Mimicking B : Change in Performance In Term 2 (Benchmark vs. $\delta = 0$)	-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]

Notes: The numbers in this table are obtained by simulating the model for 1,000,000 governors, given the structural parameters in Table 3. 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 5% level and (*) denotes significance at 10% level.

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 also argued that this coefficient will be the combination of the three effects we measured in this section, namely *discipline−pure selection+mimicking.* Table 5 shows that both discipline and pure selection are large and are highly statistically significant and that selection is reduced by mimicking, leading to the difference between discipline and selection (pure selection net of mimicking) to be statistically significant.

Table 6: **One-Term Counterfactual**

Good governors	51.8%
High Effort	51.8%
Average Performance (JAR Points)	54.0
Life-time Discounted Welfare for Voter	360.3

Note: The numbers on this table are obtained by simulating the model for 1,000,000 governors, given the structural parameters in Table 3.

5.3 Counterfactuals

We consider three 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 length, changes in cost of effort for governors, and the possibility of a signal of governor effort.

5.3.1 Term Length

The first counterfactual we consider is a change in the term limits to n terms where n is different from the currently most prevalent regime of two terms. For reasons we discuss below, we only consider $n = 1$, though we have worked out the version with $n = 3$ and a version for a generic n is a straightforward extension.[26] We choose not to pursue other values because we do not think it is reasonable to assume some of the parameter estimates remain unchanged in a regime with $n > 2$. For example the change in the term limit may change self-selection into politics and thus may alter π or an incumbent facing reelection multiple times will likely draw election shocks from a different distribution than the one she drew from in the first reelection. As such, a richer model that considers these extensions is likely required for a thorough analysis. We believe, however, that the case $n = 1$ is not a large enough change to jeopardize the validity of the structural estimates.

Table 6 reports the key results from the one-term counterfactual. Since there is no reelection, only good governors would exert high effort, leading to an average performance of 54 JAR points. Lifetime welfare for the voter is 360.3 in this case. Comparing the results in Table 6 with those in Table 4, having a two-term rather than a one-term limit is unambiguously better for the voter. First of all, more governors exert effort in their first

[26] Doing so takes some additional work – for example now governors will take in to account the outcome y_{t-1} in making their effort decision e_t in period t since the former influences elections outcomes as well.

terms, leading to a higher average JAR. This is because 27% 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 life-time welfare of the voter 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.

5.3.2 Cost of Exerting Effort

We consider two extreme exercises to highlight the importance of discipline. In the first, we set $c = 1$, in which case no bad governor would ever find it optimal to exert effort. Accordingly we label this the no-discipline counterfactual. In the second exercise, we consider the other extreme with $c = 0$, so that all bad governors choose to exert effort in their first term, and we label this the all-discipline counterfactual. Results are presented in Table 7.

In the absence of discipline, voters know that any governor who performs well is more likely to be a good governor. If there were no election shocks, most of the good governors would be reelected since their performance drawn from the high-effort distribution would imply a high probability that they exerted high effort. In that case most good governors are reelected and most bad governors lose reelection. With election shocks, more good governors would lose elections and more bad governors would win. Given the distribution of election shocks, this would imply that in the absence of discipline, 63.0% of second term governors to be of the good type. This compares with only 51.8% in the one-term case, and 59.6% in the benchmark two-term case. Since bad governors do not exert effort, there is no mimicking and the difference between 63.0% and 51.8% is the selection (and pure selection) effect. As should be clear, in this case selection is maximized. As we discussed earlier, voters are better off in the two-term-limit world relative to a one-term-limit one by about 4.2% in welfare. The decomposition in this section suggests that about 2/3 of this is due to the disciplining effect of elections: going from the one-term-limit counterfactual to the

Table 7: **No-Discipline and All-Discipline Counterfactuals**

	No-Discipline	All-Discipline
Good governors in Term 1	51.8%	51.8%
Good governors in Term 2	63.0%	51.8%
Good governors Overall	56.3%	51.8%
High effort in Term 1	51.8%	100%
High effort in Term 2	63.0%	51.8%
High effort Overall	56.3%	81.3%
Average Performance in Term 1 (JAR Points)	54.0	64.0
Average Performance in Term 2 (JAR Points)	56.3	54.0
Average Performance Overall (JAR Points)	55.0	601
Life-time Discounted Welfare for Voter	365.5	400.8

Notes: The counterfactual is 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. The solution uses the estimates of the structural parameters reported in Table 3.

no-discipline counterfactual, welfare goes up from 360.3 to 365.5 while from the no-discipline counterfactual to the benchmark model welfare goes up from 365.5 to 375.3.

In the all-discipline counterfactual, in contrast, mimicking wipes out all of pure selection as all bad governors act like good governors in the first term, and the selection effect is zero. Since all bad governors exert high effort, performance is as high as it could be in the first term, which means discipline is maximized. However since first-term performance provides no information about the type of the governors, the fraction of bad governors in the second term remains exactly at the same level as the first term, therefore it is higher relative to other versions we considered. In net, there is a substantial welfare gain relative to the benchmark case – the electorate's welfare improves by 6.8%. Welfare is still significantly lower than the first-best of 426.6 because of low performance in the second term and the partial randomness of elections. Nevertheless, these results highlight the importance of discipline for the welfare of voters.

5.3.3 Noisy Effort Signal

In our final counterfactual, we consider providing a (noisy) signal to the electorate about the effort of the governors. These will help to understand 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 8 reports discipline and selection measures (analogous to Table 5) for different values of the partially and fully

Table 8: **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%

Notes: The first column ($\zeta = 0.5$) shows the benchmark results from Tables 2, 3 and 4. Structural parameters are kept as in Table 2. See Table 4 for the definitions of the discipline and selection measures. The "B" measures are reported as percent change from the value in a world with a one-term limit. See Table 5. Welfare gain is relative to the benchmark with no signal (or equivalently with $\zeta = 0.5$).

informative signals of governor effort, the latter both in the presence and absence of an election shock. Throughout this section, we assume the structural parameters shown in Table 3 are unchanged but solve for the equilibrium objects for every probability ζ considered. We show the re-computed δ in the table.

The first column shows the benchmark results of no information 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 case of an uninformative (or no) signal, the fraction 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 8. (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 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.

We can now see why partial observability of effort implied such a small increase in the selection effect relative to the case of no observability. As we discussed in Section 5.2, the mimicking by bad governors in the first term reduces the effect of selection by making it more difficult to distinguish types based on the performance signal. Perfect observability of effort (and hence low effort making it unambiguous that a governor is bad) does not induce perfect discipline on governors when reelection has a significant exogenous random component. In the limit, when effort is perfectly observable and low effort guarantees electoral defeat, discipline is perfect (that is, there will be no governors in the third group), but the selection effect goes to zero precisely because of full mimicking by bad governors.

The last row in Table 8 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 while 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 all-discipline counterfactual we report in Section 5.3.2. 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 less 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 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 8.

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 9, 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 9 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) rather than 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

Table 9: **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 Undecideds**
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%

Notes: The top of each panel show the re-estimated 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.

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 rather than 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 9) 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, 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 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 multi-faceted measure of performance such as JAR.

6 Conclusions

In this paper 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 of 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 paper, 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 don't reelection incentives discipline a larger fraction of them? Arguing that there is a large stochastic element to elections doesn't 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 paper is a useful step in that direction.

References

[1] Alt, J., E. Bueno de Mesquita and S. Rose (2011) "Disentangling Accountability and Competence in Elections: Evidence from U.S. Term Limits," *Journal of Politics* 73(1): 171-186.

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

[3] Avis, E., C. Ferraz, and F. Finan (2016) "Do Government Audits Reduce Corruption? Estimating the Impacts of Exposing Corrupt Politicians," NBER Working Paper No. 22443.

[4] Banks, J. and R. Sundaram (1993) "Adverse Selection and Moral Hazard in a Repeated Elections Model," in W. Barnett et al. (eds.) *Political Economy: Institutions, Information, Competition and Representation*, New York, NY: Cambridge University Press.

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

[6] Besley, T. (2006) *Principled Agents? The Political Economy of Good Government*, Oxford University Press.

[7] Besley, T. and A. Case (1995) "Does Electoral Accountability Affect Economic Policy Choices? Evidence from Gubernatorial Term Limits," *Quarterly Journal of Economics* 110(3): 769-798.

[8] Besley, T. and A. Case (2003) "Political Institutions and Policy Choices: Evidence from the United States," *Journal of Economic Literature* 41(1): 7-73.

[9] Beyle, T., R. Niemi, and L. Sigelman (2002), "Gubernatorial, Senatorial, and State-level Presidential Job Approval Ratings: A Compilation of Data," *State Politics and Policy Quarterly* 2(3): 215-229.

[10] Brender, A. and A. Drazen (2005), "Political Budget Cycles in New Versus Established Democracies," *Journal of Monetary Economics* 52(7): 1271-1295.

[11] Brender, A. and A. Drazen (2008), "How Do Budget Deficits and Economic Growth Affect Reelection Prospects? – Evidence from a Large Panel of Countries," *American Economic Review* 98(5): 2203-2220.

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

[13] Dal Bó, E. and M. Rossi (2011), "Term Length and the Effort of Politicians," *Review of Economic Studies,* 78, 1237–1263.

[14] DeBacker, J. (2011), "The Price of Pork: The Seniority Trap in the US House," *Journal of Public Economics*, 95(1), 63-78.

[15] Diebold, F.X. and R. Mariano (1995), "Comparing Predictive Accuracy," *Journal of Business and Economic Statistics*, 13(3), 253-263.

[16] Diermeier, D., H. Eraslan, and A. Merlo (2003),"A Structural Model of Government Formation," *Econometrica* 71: 27-70.

[17] Drazen, A. and E.Y. Ozbay (2015), "Does 'Being Chosen to Lead' Induce Non-Selfish Behavior? Experimental Evidence on Reciprocity," working paper.

[18] Duggan, J. and C. Martinelli (2015), "Electoral Accountability and Responsive Democracy," working paper.

[19] Fearon, J. (1999), "Electoral Accountability and Control of Politicians: Selecting Good Types Versus Sanctioning Poor Performance," in Przeworski, Stokes, and Manin (eds.) *Democracy, Accountability, and* R*epresentation*, Cambridge, UK: Cambridge University Press.

[20] Ferejohn, J. (1986) "Incumbent Performance and Electoral Control," *Public Choice* 50: 5-26.

[21] Ferraz, C. and F. Finan (2011), "Electoral Accountability and Corruption in Local Governments: Evidence from Audit Reports," *American Economic Review* 101(4): 1274-1311.

[22] Finan, F. and M. Mazzocco (2016) "Electoral Incentives and the Allocation of Public Funds," NBER Working Paper 21859.

[23] Gagliarducci, S. and T. Nannicini (2013), "Do Better Paid Politicians Perform Better? Disentangling Incentives from Selection," *Journal of the European Economic Association,* 11, p. 369–398.

[24] Gowrisankaran, G., Mitchell, M.F. and A. Moro (2008), "Electoral Design and Voter Welfare from the US Senate: Evidence from a Dynamic Selection Model," *Review of Economic Dynamics,* 11(1), 1-17.

[25] Hayes, D. and S. McKee (2004), "Booting Barnes: Explaining the Historic Upset in the 2002 Georgia Gubernatorial Election," *Politics and Policy,* 32(4), 1-29.

[26] Healy, A.J., N. Malhotra, and C.H. Mo (2010) "Irrelevant Events Affect Voters' Evaluations of Government Performance," *Proceedings of the National Academy of Sciences,* 107(29), 12804-12809.

[27] Jacobson, G.C. (2006) "'The Polls': Polarized Opinion in the States: Partisan Differences in Approval Ratings of Governors, Senators, and George W. Bush," *Presidential Studies Quarterly*, 36(4), 732-757.

[28] Lahiri, K. and L. Yang (2013), "Forecasting Binary Outcomes," *Handbook of Economic Forecasting*, Vol. 2B, (eds. A. Timmermann and G. Elliott), 1025-1106.

[29] List, J. and M. Sturm (2006), "How Elections Matter: Theory and Evidence from Environmental Policy," *Quarterly Journal of Economics,* 121(4), 1249-1281.

[30] Lucas, R. (1976). "Econometric Policy Evaluation: A Critique," in *The Phillips Curve and Labor Markets*, K. Brunner and A. Meltzer eds. Carnegie-Rochester Conference Series on Public Policy 1, 19–46

[31] Maskin, E. and J. Tirole (2004), "The Politician and the Judge: Accountability in Government," *American Economic Review* 94(4): 1034-1054.

[32] Merlo, A. (1997), "Bargaining over Governments in a Stochastic Environment," *Journal of Political Economy,* 105: 101-131.

[33] Sieg, H. and C. Yoon (2016), "Estimating a Dynamic Games of Electoral Competition to Evaluate Term Limits in U.S. Gubernatorial Elections," *American Economic Review*, forthcoming.

[34] Smart, M. and D.M. Sturm (2013) "Term Limits and Electoral Accountability," *Journal of Public Economics* 107: 93-102.

[35] Strömberg, D. (2008), "How the Electoral College Influences Campaigns and Policy: The Probability of Being Florida," *American Economic Review* 98(3): 769-807.

[36] Vlaicu, R. and A. Whalley (2016) "Hierarchical Accountability in Government," *Journal of Public Economics*, 134:85-99.

[37] Wolfers, J. (2007), "Are Voters Rational? Evidence from Gubernatorial Elections," working paper.

Appendix (For Online Publication)

A Existence and Uniqueness of Equilibrium

We begin with the following lemma that higher first-term performance increases the probability the voter assigns to the incumbent being good.

Lemma 1 *The posterior probability that a governor is good, $\hat{\pi}(y_1)$, is a unique function that is monotonically increasing in y_1.*

Proof. The definition of $\hat{\pi}(y_1)$ as a posterior probability clearly implies that it is unique. To show that $\hat{\pi}$ is increasing in y_1, note that all good governors exert high effort in the first term, while bad governors exert high effort in the first term only if $c < \rho_H - \rho_L$. Since both ρ_H and $\rho_L \in (0,1)$ and since $c = 1$ for some governors, in any equilibrium some bad governors are exerting low effort in the first term. Since according to (6a) and (6b), the distribution of y_1 when $e_1 = H$ first-order stochastically dominates that of y_1 when $e_1 = L$, a higher value of y_1 must therefore raise the posterior probability that a governor is good. Formally, one can differentiate (15) with respect to y_1 to obtain

$$\frac{d}{dy_1}\hat{\pi}(y_1) = \frac{\pi(1-\pi)(1-\delta)}{2\boldsymbol{\pi}} \frac{(Y_H - Y_L)\exp\left\{-\frac{1}{2}\left[\left(\frac{y_1 - Y_H}{\sigma_y}\right)^2 + \left(\frac{y_1 - Y_L}{\sigma_y}\right)^2\right]\right\}}{\sigma_y\left\{[\pi + (1-\pi)\delta]\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) + (1-\pi)(1-\delta)\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right)\right\}^2} > 0 \tag{A-1}$$

since $Y_H > Y_L$ (where π is the prior probability that the governor is good and $\boldsymbol{\pi}$ is the mathematical constant "pi".) ■

Proof of Proposition 1. Consider first the voter's problem. The solution to the voter's problem (11) given in (12) yields a unique critical value for $\hat{\pi}(y_1)$ for any ε. This implies, given the above lemma, that there exists a unique cut-off value of y_1 for each ε, below which the voter does not reelect the incumbent, at or above which he does.

The voter's decision yields the reelection probability (16) that may be written

$$\psi(y_1) = 1 - \Phi\left[\frac{(1-\beta)\mathbb{V} - \hat{\pi}(y_1)(Y_H - Y_L) - Y_L - \mu}{\sigma_y}\right]$$

where $\Phi(.)$ denotes the CDF of a standard normal random variable. $\psi(y_1)$ is unique, so that the reelection probabilities defined by (17) and (18) are unique and contained in $(0,1)$. The lemma and $Y_H > Y_L$ further implies that $\psi(y_1)$ is monotonically increasing in y_1, so that first-order stochastic dominance implies not only that $\rho_H > \rho_L$, but also that the difference

$$\rho_H - \rho_L = \frac{1}{\sigma_y} \int \psi(y_1) \left[\phi\left(\frac{y_1 - Y_H}{\sigma_y} \right) - \phi\left(\frac{y_1 - Y_L}{\sigma_y} \right) \right] dy_1 \qquad \text{(A-2)}$$

is contained in $(0, 1)$ and is unique.

Now, consider governors. By assumption, good governors always supply high effort. The decision problem of a bad governor in (3) implies that her effort decision in the first term of office will be described by a single cutoff $c^* = \rho_H - \rho_L \in (0, 1)$, where c^* is unique given that $\rho_H - \rho_L$ is unique in equilibrium. This also implies that $\delta = \rho_H - \rho_L \in (0, 1)$ is unique. Finally, $\mathbb{V}$ – the voter's life-time expected utility one period in the future – will obviously be unique as well in the equilibrium.

Hence, taking the decision problems of voters and governors together, the equilibrium described in the paper exists and is unique.■

B Two-Term Model with a Noisy Effort Signal

We here set out some of the key equations that would differ from the unobservable effort benchmark model to complement the discussion in the text. The voter's value function, conditional on first-term observables would be:

$$W(y_1, z_1, \varepsilon) = y_1 + \beta \max_{R \in \{0,1\}} \mathbb{E} \left\{ \begin{array}{l} R\left[y_2 + \varepsilon + \beta W(y_1', z_1', \varepsilon')\right] + \\ +(1 - R) W(y_1', z_1', \varepsilon') \end{array} \middle| y_1, z_1, \varepsilon \right\} \qquad \text{(A-3)}$$

which leads to (8) with the new definition of $\mathbb{V} \equiv \mathbb{E}\left[W(y_1', z_1', \varepsilon')\right]$

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

The incumbent's posterior reputation becomes:

$$\begin{aligned}
\hat{\pi}(y_1, z_1) &\equiv \mathbb{P}(\theta = G|y_1, z_1) = \frac{p(y_1, z_1|\theta = G)\,\mathbb{P}(\theta = G)}{p(y_1, z_1|\theta = G)\,\mathbb{P}(\theta = G) + p(y_1, z_1|\theta = B)\,\mathbb{P}(\theta = B)} \\
&= \begin{cases} \frac{\pi\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)\zeta}{[\pi+(1-\pi)\delta]\phi\left(\frac{y_1-Y_H}{\sigma_y}\right)\zeta+(1-\pi)(1-\delta)\phi\left(\frac{y_1-Y_L}{\sigma_y}\right)(1-\zeta)} & \text{if } z_1 = H \\ \frac{\pi\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right)(1-\zeta)}{[\pi+(1-\pi)\delta]\phi\left(\frac{y_1-Y_H}{\sigma_y}\right)(1-\zeta)+(1-\pi)(1-\delta)\phi\left(\frac{y_1-Y_L}{\sigma_y}\right)\zeta} & \text{if } z_1 = L \end{cases}
\end{aligned}$$

because

$$\begin{aligned}
p(y_1, z_1|\theta = G) &= p(y_1, z_1|\theta = G, e_1 = H)\,\mathbb{P}(e_1 = H|\theta = G) + \\
&\quad + p(y_1, z_1|\theta = G, e_1 = L)\,\mathbb{P}(e_1 = L|\theta = G).
\end{aligned} \tag{A-4}$$

and $\hat{\pi}(y_1, z_1)$ replaces $\hat{\pi}(y_1)$ in various equations such as (10) and (14).

Reelection probabilities conditional on voter information are:

$$\begin{aligned}
\psi(y_1, z_1) &= \mathbb{P}(R = 1|y_1, z_1) = [\varepsilon > \hat{\varepsilon}(y_1, z_1)] \\
&= 1 - \Phi\left[\frac{\hat{\varepsilon}(y_1, z_1) - \mu}{\sigma_\varepsilon}\right]
\end{aligned} \tag{A-5}$$

We may then write reelection probabilities, as perceived by the incumbent:

$$\rho_H = \frac{1}{\sigma_y}\int [\zeta\psi(y_1, z_1 = H) + (1-\zeta)\psi(y_1, z_1 = L)]\,\phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) dy_1 \tag{A-6}$$

$$\rho_L = \frac{1}{\sigma_y}\int [(1-\zeta)\psi(y_1, z_1 = H) + \zeta\psi(y_1, z_1 = L)]\,\phi\left(\frac{y_1 - Y_L}{\sigma_y}\right) dy_1 \tag{A-7}$$

C Some Computational Details

We need to evaluate some integrals numerically to obtain (23) and (24) in the text. Note that all the integrals we deal with have the following general form

$$\int \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \xi(x)\, dx \tag{A-8}$$

where $x \sim N(\mu, \sigma^2)$ is a generic normal random variable and $\xi(x)$ is a known function. Let's apply a change of variables $\hat{x} = \frac{(x-\mu)}{\sqrt{2}\sigma}$, where $\hat{x} \sim N(0, 0.5)$. This also means $x = \sqrt{2}\sigma\hat{x} + \mu$.

Then, written using the pdf of $\hat{x}$ the integral simplifies to

$$\int \frac{1}{\sqrt{\pi}} \exp\left(-\hat{x}^2\right) \xi\left(\sqrt{2}\sigma\hat{x}+\mu\right) d\hat{x} \tag{A-9}$$

Finally, using Gauss-Hermite quadrature we can approximate this integral using

$$\frac{1}{\sqrt{\pi}} \sum_{i=1}^{m} \omega_i \xi\left(\sqrt{2}\sigma\hat{x}_i+\mu\right) \tag{A-10}$$

where the $\hat{x}_i$ and ω_i are the Gauss-Hermite quadrature nodes and weights respectively and m is the order of integration.

Turning to integrals in (23), they can be computed by a Gauss-Hermite approximation with $\xi_1(y_1) = \psi(y_1)$.

$$\mathcal{R}_1 \equiv \delta - \frac{1}{\sigma_y} \int \xi_1(y_1)\, \phi\left(\frac{y_1 - Y_H}{\sigma_y}\right) dy_1 + \frac{1}{\sigma_y} \int \xi_1(y_1)\, \phi\left(\frac{y_1 - Y_L}{\sigma_y}\right) dy_1 \tag{A-11}$$

As for the integrals in (24), the first one is

$$\begin{aligned} \mathbb{A}_1 &\equiv \frac{1}{\sigma_y\sigma_\varepsilon} \int\int W\left(y_1', \varepsilon'\right) \phi\left(\frac{y_1' - Y_H}{\sigma_y}\right) \phi\left(\frac{\varepsilon' - \mu}{\sigma_\varepsilon}\right) dy_1' d\varepsilon' && \text{(A-12)} \\ &= \frac{1}{\sigma_y} \int \underbrace{\left\{\frac{1}{\sigma_\varepsilon} \int W\left(y_1', \varepsilon'\right) \phi\left(\frac{\varepsilon' - \mu}{\sigma_\varepsilon}\right) d\varepsilon'\right\}}_{\xi_2\left(y_1'\right)} \phi\left(\frac{y_1' - Y_H}{\sigma_y}\right) dy_1' && \text{(A-13)} \end{aligned}$$

Using the definition of $W(.)$ in (11), $\xi_2(y_1')$ can be written as

$$\begin{aligned} \xi_2\left(y_1'\right) &= y_1' + \beta\left\{\hat{\pi}\left(y_1'\right) Y_H + \left[1 - \hat{\pi}\left(y_1'\right)\right] Y_L + \beta\mathbb{V}\right\} \frac{1}{\sigma_\varepsilon} \int_{\hat{\varepsilon}\left(y_1'\right)}^{\infty} \phi\left(\frac{\varepsilon' - \mu}{\sigma_\varepsilon}\right) d\varepsilon' \\ &\quad + \beta\frac{1}{\sigma_\varepsilon} \int_{\hat{\varepsilon}\left(y_1'\right)}^{\infty} \varepsilon' \phi\left(\frac{\varepsilon' - \mu}{\sigma_\varepsilon}\right) d\varepsilon' && \text{(A-14)} \\ &\quad + \beta\frac{1}{\sigma_\varepsilon}\mathbb{V} \int_{-\infty}^{\hat{\varepsilon}\left(y_1'\right)} \phi\left(\frac{\varepsilon' - \mu}{\sigma_\varepsilon}\right) d\varepsilon' \\ &= y_1' + \beta\left\{\hat{\pi}\left(y_1'\right) Y_H + \left[1 - \hat{\pi}\left(y_1'\right)\right] Y_L + \beta\mathbb{V}\right\} \psi\left(y_1'\right) \\ &\quad + \beta\left(\mu + \sigma_\varepsilon \frac{\phi\left(\frac{\hat{\varepsilon}\left(y_1'\right) - \mu}{\sigma_\varepsilon}\right)}{\psi\left(y_1'\right)}\right) && \text{(A-15)} \\ &\quad + \beta\mathbb{V}\left[1 - \psi\left(y_1'\right)\right] \end{aligned}$$

where for the second term we use the formula for the expected value of a truncated normal distribution

$$\mathbb{E}\left(x|x>a\right)=\mu+\sigma\frac{\phi\left(\frac{a-\mu}{\sigma}\right)}{1-\Phi\left(\frac{a-\mu}{\sigma}\right)}\text{ where }x\sim N\left(\mu,\sigma^{2}\right) \tag{A-16}$$

Thus $\mathbb{A}_1$ can be computed using a Gauss-Hermite approximation using $\xi_2\left(y_1'\right)$.

The second integral in (24) can be computed analogously

$$\begin{aligned}\mathbb{A}_2 &\equiv \frac{1}{\sigma_y\sigma_\varepsilon}\int\int W\left(y_1',\varepsilon'\right)\phi\left(\frac{y_1'-Y_L}{\sigma_y}\right)\phi\left(\frac{\varepsilon'-\mu}{\sigma_\varepsilon}\right)dy_1'd\varepsilon' \\ &= \frac{1}{\sigma_y}\int\xi_2\left(y_1'\right)\phi\left(\frac{y_1'-Y_L}{\sigma_y}\right)dy_1'\end{aligned} \tag{A-17}$$

where the only difference relative to $\mathbb{A}_1$ is that the integration is with respect to the low-effort distribution. Thus we have

$$\mathcal{R}_2\equiv\mathbb{V}-\left[\pi+\left(1-\pi\right)\delta\right]\mathbb{A}_1-\left(1-\pi\right)\left(1-\delta\right)\mathbb{A}_2 \tag{A-18}$$

D Incumbency Advantage

In Table A2 we show how the unconditional probability of winning reelections for incumbents, $\mathbb{P}\left(R=1\right)$, depends on structural parameters. Recall that we define incumbency advantage as $\mathbb{P}\left(R=1\right)>0.5$ in the text and in our sample 61.2% of incumbents win reelection. The key message of this appendix will be that incumbency advantage reflected in this number does not simply depend on μ, but on all structural parameters in complicated ways, so that $\mu>0$ is neither necessary nor sufficient in itself for the model to match the data.

In Table A2, we start by a very simple configuration of parameters: $\pi=0.5$, $\mu=0$, $\sigma_y=0$, $\sigma_\varepsilon=0$ and δ is forced to be zero by assuming very large cost of exerting effort for bad governors. In this case, all good governors exert high effort, none of the bad governors exert high effort, and since $\sigma_y=0$ effort can be perfectly observed by the voters. They reelect all good governors and replace bad governors, yielding a 50% reelection probability of incumbents. In column 2 we change π to the estimated value of 0.52 and this simply increases the reelection probability to 0.52. This shows that the fraction of good governors is important for the incumbency advantage.

In the remaining columns we allow bad governors to optimally choose their effort and thus $\delta\in(0,1)$ in all cases. Column 3 uses the same parameter configuration as column 2 where now 42% of bad governors are disciplined and the reelection probability is reduced to

30%. This is because in this equilibrium voters reelect only 42% of governors that produce a high performance (remember that effort leads deterministically to performance since $\sigma_y = 0$) and none of the governors that exert low effort and get low performance are reelected. This shows that bad governors being disciplined reduces incumbency advantage as the voters become worried about reelecting a bad governor who will exert low effort. This has a close link to the freshman effect.

In the rest of the columns we turn on the stochastic process for outcomes so that effort can no longer be inferred from performance. In column 4 now $\rho_L > 0$ since a lucky bad governor that exerts low effort can obtain a draw for y_1 that would get her reelected. Turning on the stochastic process for y increases the reelection probability since now the voters realize that they cannot be sure about the type of the governor based on performance.

In the rest of the table we show how μ and σ_ε affect the reelection probability. The last column shows the results with our benchmark parameters where the reelection probability is 63%. First in column (5) we increase μ to 5 while keeping $\sigma_\varepsilon = 0$. This gives a reelection probability that is exactly 63% – same as our benchmark specification. This means we can match the degree of incumbency advantage with $\mu = 5$ and $\sigma_\varepsilon = 0$ but, of course, this misses the observation in the data that there are a number of surprise wins and losses. In column 6 we increase σ_ε to 1 while keeping μ at 5. This reduces the reelection probability to 0.36. In columns 7 and 8 we keep increasing σ_ε towards the benchmark value of 13.07 and in order to keep the reelection probability near 63%, something estimation would aim to do since reelection outcomes are part of the observed data, we need to keep increasing μ. This means that the value of μ does not pin down the incumbency advantage but once the rest of the parameters are pinned down based on other observations, then μ adjusts to deliver the right reelection probability, and thus the right degree of incumbency advantage.

Table A1: Governors

State	Name (Year Facing Reelection)	y_1	$\hat{\pi}(y_1)$	$\psi(y_1)$	y_2	$\bar{\pi}(y_1, R, y_2)$
AL	Fob James Jr. (1982)	26.6	0%	32%	-	0%
AL	George C. Wallace (1986)	40.8	8%	37%	-	8%
AL	Guy Hunt (1990)	60.1	70%	74%	38.2	8%
AL	Fob James Jr. (1998)	48.0	28%	49%	-	28%
AL	Don Siegelman (2002)	61.2	72%	75%	-	72%
AL	Bob Riley (2006)	54.0	53%	65%	63.6	90%
AR	Mike Beebe (2010)	81.4	80%	79%	80.4	100%
CA	Pete Wilson (1994)	33.0	2%	33%	41.4	0%
CO	Bill Owens (2002)	71.0	79%	78%	57.1	89%
CO	Bill Ritter (2010)	56.9	63%	70%	-	63%
DE	Thomas R. Carper (1996)	65.8	77%	77%	80.7	100%
FL	Robert Graham (1982)	68.2	78%	78%	80.6	100%
FL	Bob Martinez (1990)	44.3	16%	41%	-	16%
FL	Lawton Chiles (1994)	38.7	6%	35%	50.0	3%
FL	Jeb Bush (2002)	66.2	77%	77%	57.9	89%
FL	Charlie Crist (2010)	66.3	77%	77%	-	77%
GA	Zell Miller (1994)	56.8	63%	70%	70.1	98%
GA	Roy Barnes (2002)	77.8	80%	79%	-	80%
GA	Sonny Perdue (2006)	61.2	72%	75%	58.1	87%
IN	Robert D. Orr (1984)	37.0	4%	34%	46.0	1%
IN	Evan Bayh (1992)	69.9	79%	78%	57.0	88%
IN	Mitch Daniels (2008)	48.7	31%	51%	65.2	84%
KS	Mike Hayden (1990)	50.0	36%	54%	-	36%
KS	Joan Finney (1994)	33.8	2%	33%	-	2%
KS	Kathleen Sebelius (2006)	63.6	75%	76%	64.9	97%
KY	Paul E. Patton (1999)	60.3	71%	74%	51.9	62%
KY	Ernie Fletcher (2007)	39.6	7%	36%	-	7%
LA	Edwin W. Edwards (1987)	21.3	0%	32%	-	0%
LA	Charles Roemer (1991)	56.9	63%	70%	-	63%
LA	Edwin W. Edwards (1995)	32.2	1%	33%	-	1%
LA	Mike Foster (1999)	77.7	80%	79%	70.4	99%

Continued on next page

Table A1 – Governors *(continued)*

State	Name (Year Facing Reelection)	y_1	$\hat{\pi}(y_1)$	$\psi(y_1)$	y_2	$\bar{\pi}(y_1, R, y_2)$
LA	Kathleen Babineaux Blanco (2007)	46.1	21%	45%	-	21%
ME	John Baldacci (2006)	51.1	41%	57%	47.3	15%
MD	Harry Hughes (1982)	32.0	1%	33%	44.8	0%
MD	William D. Schaefer (1990)	61.9	73%	76%	33.6	4%
MD	Parris N. Glendening (1998)	36.5	4%	34%	55.4	5%
MD	Robert L. Ehrlich (2006)	57.1	63%	70%	-	63%
MD	Martin O'Malley (2010)	55.3	58%	67%	57.5	76%
MI	Jennifer Granholm (2006)	52.4	46%	61%	40.8	5%
MS	Ray Mabus (1991)	58.6	67%	72%	-	67%
MS	Kirk Fordice (1995)	50.1	37%	55%	53.8	38%
MS	Ronnie Musgrove (2003)	68.0	78%	78%	-	78%
MS	Haley Barbour (2007)	54.1	54%	65%	67.4	96%
MO	Bob Holden (2004)	40.2	8%	36%	-	8%
MO	Matt Blunt (2008)	41.3	9%	37%	-	9%
MT	Marc Racicot (1996)	77.4	80%	79%	79.9	100%
MT	Judy Martz (2004)	29.4	1%	33%	-	1%
MT	Brian Schweitzer (2008)	72.3	79%	78%	66.7	98%
NE	Bob Kerrey (1986)	84.0	80%	79%	-	80%
NE	Kay A. Orr (1990)	52.7	48%	61%	-	48%
NE	Ben Nelson (1994)	63.9	75%	77%	77.8	100%
NV	Kenny C. Guinn (2002)	59.7	70%	74%	62.6	94%
NV	Jim Gibbons (2010)	31.8	1%	33%	-	1%
NJ	Richard J. Hughes (1965)	58.1	66%	72%	64.1	95%
NJ	Thomas H. Kean (1985)	65.6	77%	77%	72.7	99%
NJ	James J. Florio (1993)	30.0	1%	33%	-	1%
NJ	Jon Corzine (2009)	48.1	29%	50%	-	29%
NM	Gary E. Johnson (1998)	51.2	42%	58%	53.7	42%
NM	Bill Richardson (2006)	62.2	74%	76%	55.3	80%
NC	James G. Martin (1988)	65.6	77%	77%	55.2	82%
NC	James B. Hunt Jr. (1996)	68.6	78%	78%	69.9	99%
NC	Mike Easley (2004)	56.9	63%	70%	59.3	85%
NC	Bev Perdue (2012)	44.3	16%	41%	-	16%

Continued on next page

Table A1 – Governors *(continued)*

State	Name (Year Facing Reelection)	y_1	$\hat{\pi}(y_1)$	$\psi(y_1)$	y_2	$\bar{\pi}(y_1, R, y_2)$
OH	Richard F. Celeste (1986)	55.9	60%	68%	63.6	93%
OH	George V. Voinovich (1994)	64.3	76%	77%	74.7	100%
OH	Bob Taft (2002)	83.9	80%	79%	20.3	0%
OH	Ted Strickland (2010)	59.6	69%	74%	-	69%
OK	Henry L. Bellmon (1990)	45.8	20%	44%	-	20%
OK	David Walters (1994)	40.8	8%	37%	-	8%
OK	Frank Keating (1998)	68.2	78%	78%	60.5	94%
OK	Brad Henry (2006)	69.0	78%	78%	71.0	99%
OR	Barbara Roberts (1994)	28.6	1%	32%	-	1%
OR	John Kitzhaber (1998)	61.8	73%	75%	59.8	91%
OR	Ted Kulongoski (2006)	47.2	25%	47%	46.0	6%
PA	Richard L. Thornburgh (1982)	49.6	35%	53%	44.8	8%
PA	Robert P. Casey (1990)	81.3	80%	79%	33.2	5%
PA	Ed Rendell (2006)	55.0	57%	67%	52.3	50%
RI	Lincoln C. Almond (1998)	44.5	17%	42%	55.7	23%
RI	Donald Carcieri (2006)	59.4	69%	73%	46.4	32%
SC	Carroll Campbell (1990)	73.3	79%	78%	70.6	99%
SC	David Beasley (1998)	60.8	71%	75%	-	71%
SC	Jim Hodges (2002)	46.7	23%	46%	-	23%
SC	Mark Sanford (2006)	54.9	56%	66%	47.4	25%
SD	M. Michael Rounds (2006)	75.0	80%	79%	65.1	98%
TN	Ned R. McWherter (1990)	70.3	79%	78%	51.4	70%
TN	Don Sundquist (1998)	61.6	73%	75%	47.7	43%
TN	Phil Bredesen (2006)	61.8	73%	75%	74.2	100%
WV	Arch A. Moore Jr. (1988)	42.7	12%	39%	-	12%
WV	Gaston Caperton (1992)	28.3	1%	32%	36.7	0%
WV	Cecil H. Underwood (2000)	59.7	70%	74%	-	70%
WV	Bob Wise (2004)	58.9	68%	73%	-	68%
WY	Jim Geringer (1998)	66.3	77%	77%	50.0	61%
WY	Dave Freudenthal (2006)	73.4	79%	78%	78.0	100%

Notes: y_1 and y_2, if the governor is reelected, show the JAR performance of the governor. $\hat{\pi}(y_1)$ is the updated probability of the governor being good, and $\psi(y_1)$ is the probability that the governor will win re-election, both conditional on first-term performance. $\bar{\pi}(y_1, R, y_2)$ is the probability that the governor is good, having observed both terms' performance, where available.

Table A2: Incumbency Advantage and Structural Parameters

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	Benchmark
Parameters								
π	0.50	0.52	0.52	0.52	0.52	0.52	0.52	0.52
σ_y	0.00	0.00	0.00	9.84	9.84	9.84	9.84	9.84
μ	0.00	0.00	0.00	0.00	5.00	5.00	22.00	25.53
σ_ϵ	0.00	0.00	0.00	0.00	0.00	1.00	10.00	13.07
Equilibrium Objects								
δ	0	0	0.42	0.38	0.53	0.48	0.33	0.27
ρ_L	-	-	0.00	0.01	0.23	0.03	0.40	0.45
ρ_H	-	-	0.42	0.39	0.77	0.51	0.73	0.72
$\mathbb{P}(R=1)$	0.50	0.52	0.31	0.36	0.63	0.36	0.63	0.63

Notes: The parameters not shown are kept fixed at their values in Table 2. In columns (1) and (2) δ is forced to be 0, which means the bad governors are assumed to have high enough cost of exerting effort so that none of them choose to exert high effort.