Legislative Behavior Absent Reelection Incentives: Findings from a Natural Experiment in the Arkansas Senate*

Rocío Titiunik †
Department of Political Science
University of Michigan

Andrew Feher‡
Research Analyst
Covered California

March 27, 2017

Abstract

We analyze the impact of removing reelection incentives on the individual legislative participation of state lawmakers using an original study based on the random assignment of term length that occurs in the Arkansas Senate, which in turn induces the random assignment of the total number of terms each senator may serve. Across five measures of legislative output—bills introduced, bills passed, bills cosponsored, resolutions, and abstention rates, we are unable to reject the null hypothesis of no effect. Since our sample is small, we adopt two strategies in our statistical analysis: we perform randomization-based inference to ensure that our tests adequately control size, and we use tests of equivalence to avoid incorrectly concluding that the effects are null because of low power. We also use bounds as a robustness check to address attrition in our original experimental sample.

Keywords: Term limits; Legislative behavior; Fisherian inference; Tests of equivalence; Bounds.

*We thank Steve Cook, the Chief Council of the Arkansas State Senate, for providing details regarding the drawing of lots following censuses; Blake Potts at the Arkansas Legislative Digest for providing the roll-call data; the Roy Pierce Fellowship at the Center for Political Studies, University of Michigan, for financial support; and the editor, Linda Sharples, the associate editor, two anonymous reviewers, Matias Cattaneo, Jeremy Gelman, Marjorie Sarbaugh-Thompson, Chuck Shipan, and participants at the 2014 State Politics and Policy Conference, for thoughtful suggestions and discussions that greatly improved our manuscript.

†James Orin Murfin Associate Professor, Department of Political Science, University of Michigan, <titiunik@umich.edu>, http://www.umich.edu/~titiunik, Department of Political Science, University of Michigan, 5700 Haven Hall, 505 South State St., Ann Arbor, MI 48109.

‡Research Analyst, Covered California, <andrew.feher@covered.ca.gov>.

1 Introduction

In 1990, voters in California, Colorado and Oklahoma approved initiatives limiting the number of terms that a lawmaker could serve in the state legislature, setting in motion one of the most significant policy changes in state government in decades. Between 1992-1996, seventeen more states followed suit. As of April 2015, fifteen states jointly housing 37% of the total U.S. population still limit state legislators' length of service. The consequences of these reelection restrictions on the quality of representation can be in principle positive or negative, as the benefits of high turnover induced by term limits may induce high costs if removing reelection incentives leads legislators to, for example, decrease their effort or adopt 'out of step' ideological positions.

We study the impact of removing reelection incentives via the adoption of term limits on state lawmakers' individual legislative output and participation. Establishing empirically whether these effects exist and measuring their magnitude is important for several reasons. First, we wish to understand whether rules that allow legislators to serve without the prospect of future electoral accountability result in systematic changes in legislative output. Since term limits are still being adopted and modified in many states, their effects on legislative behavior should be incorporated in the future design and evaluation of these policies. Second, the immediate effect of term limits policies on the legislative output of individual lawmakers may have downstream consequences for the way in which policy is produced in the states. State governments have significant authority to formulate and implement policy in areas as varied as environmental protection, intrastate commerce, and education (Gerber and Teske, 2000), and state legislatures can have an impact on the nature of policy adoption and diffusion (Shipan and Volden, 2006; Nicholson-Crotty, 2009).

Our study is based on original data from a natural experiment in the Arkansas Senate. The Arkansas Constitution includes two features—the random assignment of senators' term length in the first election after reapportionment, and term limits—the combination of which results in the random assignment of state senators to shorter or longer time horizons. In particular, senators randomly assigned four-year terms in the first election after reapportionment see one less session in office than those randomly assigned two-year terms, which allows us to examine two legislative sessions in 1997 and 2007 where the group of legislators assigned four-year lots is ineligible for reelection while the group assigned two-year lots is still eligible for one last reelection. Since this

natural experiment is based on an initial random assignment, it has the advantage of clearly defining which groups should be compared to make inferences (Sekhon and Titiunik, 2012). Drawing on these experiments, we empirically test hypotheses regarding the effects of removing immediate reelection incentives on several measures of individual legislative participation and output.

Although the effects we study are informative about a crucial aspect of term limits—the removal of reelection incentives—our design does not directly manipulate term limits and thus cannot capture the overall effect of adopting term limits in a legislature (in fact, all senators in our sample are term-limited). The ideal experiment to study term limits effects would randomly assign state legislatures to adopt either term limits or indefinite reelection, and would compare legislature-level outcomes from each group. Such an experiment would reveal the overall effect of adopting term limits, which would include both the effect of term limits on the composition of the legislature (e.g., term limits may discourage certain types of candidates from entering the race), and the effect that the removal of reelection incentives would cause on the behavior of individual legislators. Our natural experiment stands in contrast to this ideal experiment, as it manipulates how many terms a senator can serve, in a given legislature, before becoming ineligible to run for reelection. Our treatment group comprises senators who are serving their last term and are not eligible for reelection when the term they are serving expires; in contrast, our control group comprises senators who are eligible for one last reelection at the end of the term they are serving. The comparison of these groups captures the effect of removing reelection incentives relative to a condition where reelection incentives are in place for one additional electoral cycle. Our natural experimental design is thus informative about reelection incentive effects, but not about legislature-level effects that would require a design akin to the ideal experiment.

Our data has two features that result in considerable statistical challenges: small sample size and sample attrition. Our empirical analysis employs different tools to address these challenges, illustrating how such tools can be employed in other applications facing similar obstacles. The small sample size occurs because the Arkansas Senate has only thirty-five members. Although our analysis pools two cohorts, the total sample size in our experiment is only 64, because our research design forces us to discard some members. The sample attrition stems from the electoral defeat or retirement of some senators in the intervening years between their first election to count against term limits, and their last (or second to last) term.

In many cases, small sample sizes lead to two problems—hypothesis tests based on large-sample approximations have size different from their nominal level (or "incorrect size" for brevity), and the power of tests to detect departures from the null hypothesis is low. To address concerns about size, we use Fisherian randomization-based inference techniques, which are exact in finite samples. In the Fisherian randomization inference framework, the distribution of the test statistic under the sharp null hypothesis of no effect is entirely determined by the distribution of treatment assignment, which allows us to test the hypothesis of no effect for any unit with an exact finite-sample p-value instead of relying on large-sample approximations that may be inadequate in small samples. The seminal ideas are due to Fisher (1935); for a contemporaneous overview, see Imbens and Rubin (2015) and Rosenbaum (2002b).

Although a Fisherian approach leads to inferences of correct size, it does not solve the problem of low statistical power induced by our small sample size. We employ two strategies to address low power concerns. First, our choice of test statistic in our randomization-based tests is guided by power considerations; we employ various statistics tailored to detecting different types of departures from the sharp null hypothesis. Second, we employ randomization-based tests of equivalence to be able to rule out large effects. Since in our application we are unable to reject the null hypothesis of no last-term effects and we are interested in asserting that the absence of reelection incentives does not alter individual legislative output, we test the null hypothesis that the outcomes of reelection-ineligible lawmakers *differ* from those of reelection-eligible lawmakers.

In tests of equivalence, which are common in biomedical studies but rarely used in the social sciences, the type I error rate is the probability of declaring that the two groups compared are equivalent when in fact they are not (Berger et al., 1996). Thus, when controling the type I error rate, we control the probability of declaring that removing reelection incentives has no effect when in fact it does. Following related ideas developed by Rosenbaum (2010), we adapt this procedure to a randomization-based framework based on a test of the sharp null hypothesis under a constant treatment effect model, and use it to calculate, for every outcome, the minimum discrepancy between the reelection-ineligible and reelection-eligible group that leads to a rejection of the null hypothesis that the effect of removing reelection is non-zero. As we show, these tests of equivalence allow us to assert with 95% confidence that the removal of reelection incentives induced by term limits does not have very large negative impacts on most measures of individual legislative output

and participation; however, we cannot rule out moderate or small negative effects.

Our analysis reveals that, in low power settings, equivalence tests are most informative when two conditions hold: (i) theoretical expectations are one-sided, and (ii) observed point estimates or test statistics run counter to those expectations. When both conditions hold, tests of equivalence offer researchers the ability to rule out a large proportion of effects anticipated by theoretical expectations, leading to meaningful conclusions even with a small sample. As we discuss and show below, both conditions hold in our application. The literature on last-term effects emphasizes that removing reelection incentives may induce legislators to exert low effort or "shirk," leading to the expectation that removing reelection will lead to lower legislative output. Moreover, for all except one of the individual legislative outcomes we analyze, the average output in the reelection-ineligible group is higher than in the reelection-eligible group. This allows us to rule out large negative effects based on a constant treatment effect model employing test statistics that measure location shifts. Our analysis also shows that, in the absence of one or both of the above conditions, the usefulness of equivalence tests to draw conclusions from very small samples is more limited. For example, if theoretical expectations are two-sided, ruling out effects in one direction can still leave researchers with considerable uncertainty. Or, if the theoretical expectations are one-sided and the observed value of the test statistic is consistent with those expectations, the ability to reject the null hypothesis and assert that the effects are null or small is reduced. Our analysis of the number of bills cosponsored by each legislator illustrates this phenomenon: since reelection-ineligible senators cosponsor on average fewer bills than their reelection-eligible counterparts, large negative last-term effects on cosponsorship cannot be ruled out with our small data.

Finally, we address the attrition in our experimental sample using the partial identification framework developed by Manski (2003, 2007). We shift our focus to the average treatment effect of removing reelection incentives, and estimate bounds on this effect under the assumption that those lawmakers whose outcomes we do not get to observe would have exhibited systematically high or low productivity and participation, potentially affecting our conclusions. This analysis shows that non-random attrition does not seem to be driving most of our conclusions. However, it also shows that if the most extreme systematic differences between senators who survive and senators who are defeated were true, our conclusion that removing reelection incentives does not induce large negative effects would need to be moderated.

The remainder of the paper is organized as follows. In the next section, we develop theoretical expectations about last-term effects on legislative behavior in state legislatures. We then present the details of our experimental research design, followed by a section that discusses randomization inference and tests of equivalence. Next, we present our results, including a subsection with robustness checks based on bounds. We conclude in the last section. Additional results are presented in an online Supplemental Appendix.

2 Theoretical Expectations: Legislative Behavior Absent Reelection Incentives

Under an accountability model of representation, voters incorporate politicians' past actions into their voting decisions and elections are a mechanism to sanction representatives' behavior. In turn, the threat of punishment induces reelection-seeking politicians to behave in accordance with constituents' preferences and expectations. Under this model, the logical consequence of removing the possibility of reelection is to induce undesirable legislative behavior or shirking, as the threat of punishment is removed and legislators have no incentive to please the electorate (see Mansbridge, 2009, and references therein). Thus, if elections' main role is to serve as an accountability mechanism, removing the possibility of running for reelection should result in systematic changes in legislative behavior (Fearon, 1999, p. 63).

Moreover, there may be mechanisms by which the removal of reelection incentives may result in lower legislative participation and output that are not directly related to the removal of electoral accountability. One such mechanism is the potential opportunity costs of seeking future employment. If legislators harbor some degree of progressive ambition and hope to secure their next occupation before their tenure comes to an end, lame-duck legislators face a trade-off: continue to actively participate in legislative activities or curb some of that legislative participation in order to invest attention in surveying their future employment options. Given the time commitments associated with casework and with building the coalitions needed to successfully navigate bills through the legislature, we might expect those who are in their last term, and thus more likely to be in search of their next job, to reduce the effort they expend on constituency service and policymaking.

These two non-exclusive scenarios—shirking induced by the removal of reelection incentives and

shirking induced by the opportunity costs of securing future employment—imply that reelection-ineligible legislators will have less time or incentives to meet with staff to help draft legislation, to learn about the kinds of bills their colleagues are sponsoring, to jockey for support in committee and on the floor to ensure passage of the bills that they do introduce, to allocate attention to casework, and to attend roll-call votes. Accordingly, we can hypothesize that at the level of the individual legislator, reelection-ineligible members of the chamber will reduce their effort, introducing or cosponsoring fewer bills, achieving passage of fewer bills, performing less constituency service, and abstaining on a greater proportion of roll-call votes than reelection-eligible members.

Last-term effects may also depend on the degree of the legislature's professionalism. Facing low pay, limited staff resources, poor advancement prospects, and the absence of a reelection incentive, members serving in one of the twenty-four states with low-professionalization legislatures have few incentives to actively participate. By contrast, in professional legislatures, not only do staff subsidize the cost of policymaking, but these legislators earn salaries that permit them to devote all of their time to legislating. The effects of removing electoral accountability might therefore be amplified in less professional legislatures.

The Arkansas State Legislature is not highly professionalized. In Squire's (2007) index of legislative professionalism, the Arkansas Legislature ranked 39th in 1996 and 41st in 2003, and in the National Conference of State Legislature's (NCSL) Red-White-Blue trifurcation, Arkansas is considered a "White," or hybrid, legislature based on its intermediate-sized staff and salary, as legislators do not earn enough to make a living without having other sources of income (National Conference of State Legislatures, 2009). The General Assembly holds its regular session in odd-numbered years, meeting for approximately sixty days, and holds what are variously known as fiscal sessions or extraordinary sessions in even-numbered years.

During the 1980s, before term limits were adopted, legislative turnover in Arkansas was low: approximately half of the house seats and two-thirds of the senate seats were occupied by veteran legislators during this decade (Sarbaugh-Thompson, 2010, Table 1). Given this low turnover before term limits and the stringent limits on length of service that followed, the effects of term limits on the Arkansas Senate might be starker than in most other states, where the institutional change induced by more lenient term limit policies did not represent such a drastic change (Sarbaugh-Thompson, 2010, p. 202).

On the other hand, there are countervailing forces that may attenuate the potential effects of removing reelection incentives. First, some term limits advocates expected that by introducing term limits, lawmakers would be less preoccupied with reelection-centered activities, and would expend more effort on policymaking and related legislative activities (Will, 1992; Glazer and Wattenberg, 1996). If this argument is true, the extra available time enjoyed by reelection-ineligible legislators should more than compensate for the incentives to expend less effort. Second, most senators in our sample are elected or reelected in very lopsided elections (the average vote share in our sample is 88%), suggesting that reelection pressures will be lower than in other more competitive settings. Third, the need to secure future employment is likely not a major factor in part-time legislatures, implying that the opportunity cost of seeking future employment is not likely to induce additional negative effects on legislative output. Finally, if most state senators ran for higher office after serving their last term in the Senate, we would expect the future higher-office election to act as a strong incentive even when reelection is no longer possible in the current office. This point, however, does not seem to be very relevant in the Arkansas context, where only 10 of the 64 senators in our sample sought higher office after serving in the Arkansas Senate. In particular, four senators ran for U.S. House seats (Steve Bryles, Gene Jeffress, Jay Bradford, Vic Snyder), four ran for U.S. Senate seats (Gilbert Baker, Kim Hendren, Jim Holt, Lu Hardin), and three ran for Governor (Shane Broadway, Jim Holt, Mike Beebe)—see CQ Press (2016).

These countervailing forces may lead to the expectation that removing reelection incentives in the Arkansas Senate will have null or positive effects on legislative output. For this reason, all the initial statistical tests we present are two-sided. At the same time, we are particularly interested in the hypothesis that the removal of reelection incentives has negative implications for the effort and legislative output of individual members, both because we wish to rule out—or show that we cannot rule out—some of the potential negative normative implications of removing reelection incentives, and because previous non-experimental studies of term limits in state legislatures in the United States have mostly focused on (and presented evidence for) the one-sided theoretical expectation of shirking. For example, Carey et al. (2006) and Powell et al. (2007) find that term-limited state legislators devote less time to securing rewards for their district and helping constituents deal with government. In addition, Sarbaugh-Thompson et al. (2004) report that term-limited legislators turn their attention away from constituents and toward interest groups. As we discuss in the

introduction, the use of equivalence tests is most helpful to asses this one-sided hypothesis with our limited sample size.

Finally, we emphasize that it is crucial to interpret our empirical results in the Arkansas context: even if the individual legislative output of Arkansas senators whose reelection incentives are removed is not smaller than the output of their reelection-eligible peers who face the incentives of one last upcoming election, the same policy might produce negative effects in a context where, for example, members were full-time legislators and electoral competition were high.

3 A Research Design Based on Random Assignment

A possible strategy to study the effect of removing reelection incentives is to compare the outcomes of legislators who are serving their last term to the outcomes of those legislators who can still run for reelection. This type of observational research design is a common choice (see, e.g., Carey et al., 1998, 2006; Powell et al., 2007; Sarbaugh-Thompson et al., 2004), and it is often the only one available. The challenge is that politicians serving their last term in office are often systematically different from those whose electoral horizons are longer, which complicates the ability to attribute last-term behavior to the lack of electoral incentives. This phenomenon is most evident when the decision to retire is entirely under the control of the individual politicians—e.g., legislators may retire preemptively when they anticipate that their past performance may result in a future loss. But these inferential complications do not necessarily disappear when the occurrence of the last term is determined by an exogenous rule such as term limits.

In order to address some of these methodological challenges, we use a natural experiment that relies on the random assignment of term length, which in turn induces the random assignment of senators to a different maximum number of terms allowed in office. This research design avoids important inferential challenges since the group of reelection-ineligible and reelection-eligible legislators are on average identical at baseline due to the initial random assignment.

Our research design is based on the random assignment of term length in the Arkansas Senate. Arkansas senators normally serve a term of four years and their terms are staggered, with (roughly) half of the 35 senate seats up for election every two years. However, Article 8, Section 6 of the state's constitution mandates that, in the first election following a decennial census and the corresponding

redrawing of district boundaries, all 35 seats must be up for election. Since the simultaneous election of all 35 seats breaks the staggering of terms, term lengths are randomly assigned to return the chamber to staggered terms.

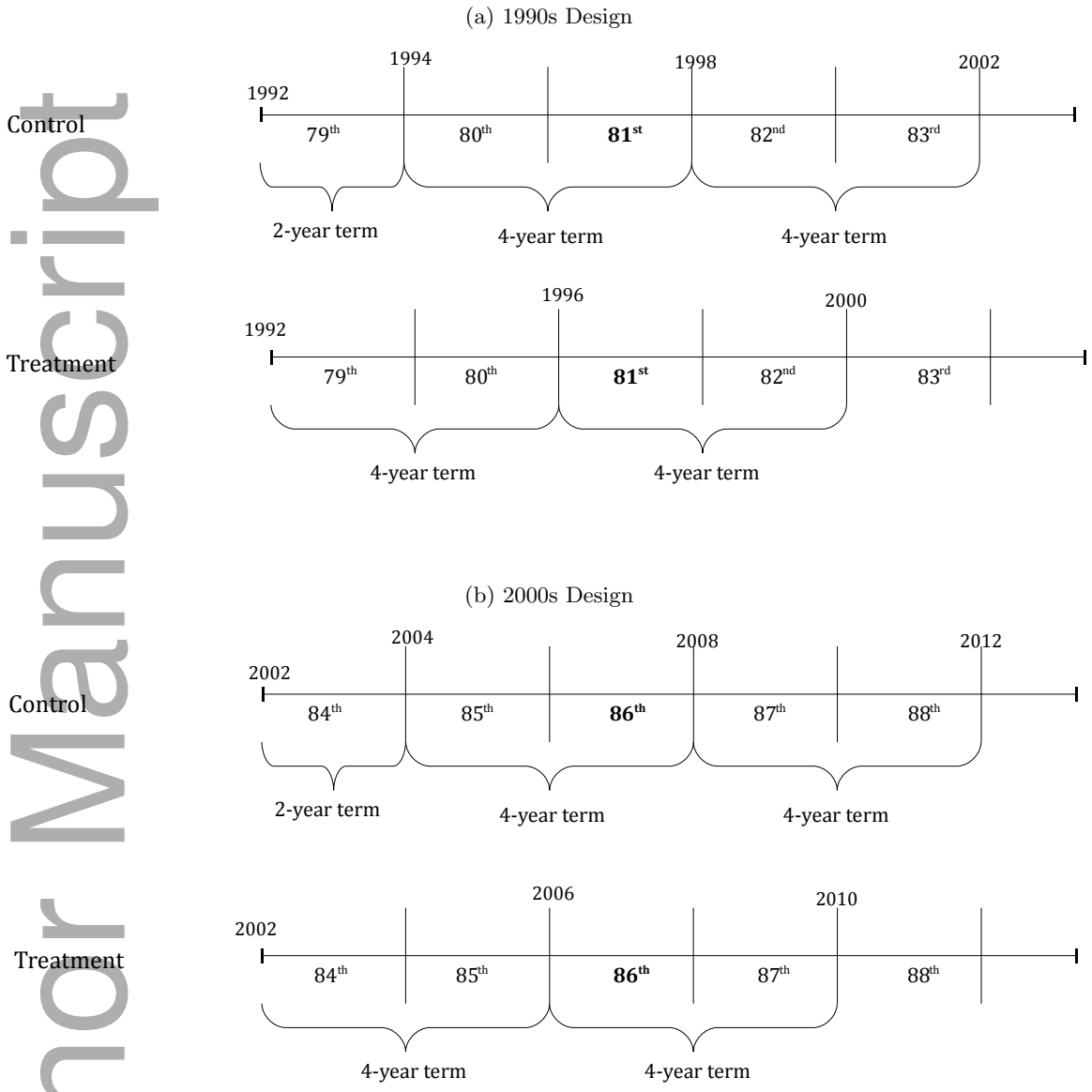
Specifically, Section 6, Amendment 23, of the Arkansas Constitution instructs senate seats to be randomly divided into two classes of size 17 and 18 after each reapportionment. The pattern of term length differs by class: senators elected to a seat in the class of size 18 serve a two-year term immediately following reapportionment and a four-year term thereafter, while senators elected to a seat in the other class serve two successive four-year terms immediately following redistricting and a two-year term at the end of the decade. Senators draw lots at the beginning of the first legislative session immediately after redistricting to determine the composition of each class of seats. This design, and similar designs in Illinois and Texas, was used by Titiunik (2016) to study the effects of term length on legislative behavior. We confirmed that the randomization procedure took place in phone conversations with the Arkansas Senate. The terms were assigned by drawing plastic eggs containing a piece of paper annotated with the number 2 or 4 from a jar, with a total of 18 eggs in the two-year term category and 17 eggs in the four-year term category, ensuring a fixed-margins randomization.

In November 1992, 60 percent of Arkansas voters supported Amendment 73, a term limits initiative that was among the most stringent in the country. This amendment limited state representatives' service to a lifetime maximum of three two-year terms and state senators' service to a lifetime maximum of two four-year terms. An important element of our research design is that two-year terms do not count against the two-term limit for senators—only four-year terms do.

Our goal is to measure outcomes for reelection-ineligible and reelection-eligible senators during the same legislative session, to avoid conflating time differences and genuine last-term effects. For this reason, we study two cohorts of senators for whom term limits become effective during the same legislative session: those first elected or reelected in 1992, and those first elected in 2000 or 2002. Figures 1(a) and 1(b) illustrate the sequence senators experience based on whether they draw a two-year or a four-year term in the 1990s and 2000s, respectively. As an example, consider two senators who are elected in November 2002, one assigned a two-year term and one assigned a four-year term. Because a two-year term does not count toward the two-term lifetime limit, the senator assigned to serve a two-year term will stand for reelection in November 2004 and again

in November 2008. By contrast, a state senator assigned a four-year lot in 2002 is already on the term-limit clock and will only stand for reelection one time in November 2006. In turn, this makes for a legislative session in 2007 (the 86th) where senators assigned four-year lots in 2002 are reelection ineligible while senators assigned two-year lots in 2002 still face an election in 2008. The sequence is analogous in the 1990s; however, due to initial confusion surrounding the passage of Amendment 73, lots were drawn after the 79th session in October 1993 instead of at the beginning of the session. In the post-2000 reapportionment, by contrast, lots were drawn in December 2002, after the election but before the start of the legislative session.

Figure 1: Illustration of Research Design



Note: The numbers in bold (81st and 86th) refer to the legislative sessions in which we examine our outcomes.

For analysis, we pool both cohorts, including all senators elected in 1992—totaling 35—and all senators elected for the first time in 2000 or 2002—totaling 29 (below we explain why we discard 6 senators). All senators in our treatment group are randomly assigned four-year terms following the 1990 or 2000 reapportionment and are thus ineligible to run for reelection after the 82nd or 87th Legislative Sessions. In contrast, all senators in our control group are assigned two-year terms after the 1990 or 2000 reapportionment, and thus are eligible for one more reelection at the end of

the 81st or 86th Legislative Sessions. We study outcomes related to legislative output during the 81st and 86th regular sessions of the Arkansas General Assembly (marked in bold in Figure 1).

Although there are some new senators elected for the first time in the middle of the 1990 and 2000 decades, we focus on the 1992 and 2002 cohorts because they define a group of senators who are all term limited at the same time—except for the two-year difference induced by the initial random assignment. As illustrated in Figure 1(a), all Arkansas senators elected in November 1992, whether elected for the first time or reelected, served their last period either in 1996-2000 or 1998-2002 (if they did not retire or lose sooner). Throughout, we refer to the length of terms by an interval from the year when the election took place to the last year of the senator's term—for example, 1998-2002 refers to the term, served between January 1999 and December 2002, for which a senator was elected in November 1998. Since 1992 is the "baseline" year when term limits are adopted, regardless of how many times senators in this cohort had been reelected prior to 1992, they would all serve their last allowed term at the same time, except for the 2-year discrepancy induced by the staggering (barring retirements and defeats).

The situation for later cohorts is different, because as some senators lose or retire before the maximum allowed number of terms, the newly elected senators' last allowed terms occur at different points in time. If a few new senators were entering every year, it would be hard to study an additional cohort, as everyone would be term-limited at different times, complicating our design. Luckily, there are only six senators who are elected for the first time between 1994 and 1998, with the remaining 29 first elected in either 2000 or 2002. These 29 senators constitute the second cohort in our analysis.

4 Improving Statistical Inferences When Samples Are Small

The sample size in our experimental study is only 64, raising the concern that standard statistical inference techniques may be inadequate for at least two reasons. First, asymptotic distributions may provide very poor approximations to the finite-sample null distribution of the relevant test statistics, leading to tests whose size may differ from their nominal level. Second, the ability to detect departures from the null hypothesis may be limited by low statistical power. We employ two strategies for statistical inference that allow us to address these challenges. To address concerns

about size, we employ Fisherian randomization-based inference, which leads to tests that are finite-sample exact. To address concerns about power, we employ two strategies: we use covariate-adjustment and other appropriate choices of test statistics in a randomization-based framework, and build randomization-based tests of equivalence under a constant treatment effect model where the null hypothesis is that the treatment has a non-zero effect.

We adopt the potential outcomes framework and let y_{1i} and y_{0i} be the potential outcomes of interest for legislator i under the reelection-ineligible state and the reelection-eligible state, respectively, for $i = 1, 2, \ldots n$. Pooling both cohorts we have a total of $n = 64$ senators. The treatment indicator is T_i, with $T_i = 1$ if senator i is reelection-ineligible and $T_i = 0$ if senator i is reelection-eligible, and n_1 and $n_0 = n - n_1$ denote the number of treated and control senators, respectively. The observed outcome or response is therefore $Y_i = T_i y_{1i} + (1 - T_i) y_{0i}$, where we follow the convention of employing lower case letters to denote fixed variables, and upper case letters to denote random variables. We collect the n observed responses in the vector $\mathbf{Y}$, and the n individual treatment assignments in the vector $\mathbf{T}$. The initial randomization to different term lengths, which in turn determines whether a senator is reelection-ineligible or not in 1997 or 2007, ensures that the distribution of the treatment T_i is not a function of the potential outcomes, so that, in the absence of complications, comparing reelection-ineligible and reelection-eligible senators is a valid strategy to learn about the effect of removing reelection incentives.

Addressing Concerns About Size: Randomization-based Inferences

The randomization-based inference framework was first introduced by Fisher (1935), and has been recently used in natural experiments and observational studies (e.g. Bowers et al., 2013; Cattaneo et al., 2015; Ho and Imai, 2006; Imbens and Rosenbaum, 2005). For an introduction, see Rosenbaum (2010, §2) and Imbens and Rubin (2015, §5); a more advanced treatment can be found in Rosenbaum (2002b). We now briefly review the most essential aspects of this framework.

In the Fisherian framework, the potential outcomes are seen as fixed and the only randomness in the model stems from the randomization of the treatment assignment. As is common, we explicitly incorporate this in our notation, using upper case for the treatment and the observed outcome, (Y_i, T_i), but lower case for the potential outcomes, (y_{1i}, y_{0i}). The most attractive methodological feature of this setup is that, given knowledge of the randomization distribution of the treatment

assignment, the so-called sharp null hypothesis of no treatment effect—$H_0 : y_{i1} = y_{i0}$ for $i = 1, 2, \ldots, n$—can be tested with no additional assumptions (Rosenbaum, 2002b), even in cases where there is interference between units, as we discuss briefly below.

When the randomization procedure is known, we can define the set Ω of all possible values of the vector $\mathbf{T}$ in which the number of treated subjects is fixed to be n_1. In the randomization of term lengths that occurs in Arkansas every decade after reapportionment, the number of elements in the set Ω is all possible values of the vector $\mathbf{T}$ in which there are $n_1 = 17$ ones and $n_0 = n - n_1 = 35 - 17 = 18$ zeros. Each of these possible assignments has an equal probability of occurring, $\mathbb{P}(\mathbf{T} = \mathbf{t}) = 1/\binom{n}{n_1}$.

To test the sharp null hypothesis H_0, we define a test-statistic $W(\mathbf{T}, \mathbf{Y})$ which depends on the treatment assignment $\mathbf{T}$ and the vector of outcomes $\mathbf{Y}$. Since the potential outcomes are assumed fixed, under H_0 the *only* random variable is the treatment assignment, implying that the distribution of $W(\mathbf{T}, \mathbf{Y})$ is completely determined by the randomization distribution of $\mathbf{T}$. The two-sided significance level for a test that rejects of H_0 is given by

$$p = \frac{\#\{\mathbf{T} \in \Omega : |W(\mathbf{T}, \mathbf{y})| \geq |W(\mathbf{t}, \mathbf{y})|\}}{\binom{n}{n_t}},$$

where $W(\mathbf{t}, \mathbf{y})$ is the observed value of the test statistic and $\#\{\cdot\}$ denotes the number of elements in a set.

If n were sufficiently small, this p-value could be calculated exactly. In our experiment, however, the enumeration of all possible values of the test statistic is unfeasible. We thus base our tests on 10,000 simulations, each simulation taking one treatment assignment at random from all possible assignments. Since the random assignment of terms was done separately for each cohort, in our simulations we separately draw the treatment assignment of each cohort, and then pool both cohorts to compute the difference in means between our pooled treatment and control groups.

The Fisherian randomization-based framework can also be used to test other hypotheses in addition to the sharp null hypothesis, and to invert such tests to obtain confidence intervals for treatment effect parameters. However, there is an important asymmetry. Unlike the sharp null hypothesis, which can be tested with no assumptions other than knowledge of the treatment randomization distribution, the construction of confidence intervals and equivalence tests requires a

treatment effect model.

The crucial feature that enables randomization-based tests to be finite-sample exact is knowledge of the exact distribution of test statistics under the null hypothesis. In turn, knowledge of the null distribution depends on the ability to impute both the treated and control potential outcomes for every unit. This is straightforward when we are testing the hypothesis that $y_{i0} = y_{i1}$ for all i: under this sharp null hypothesis, all potential outcomes are known. The challenge with testing other null hypotheses using a Fisherian framework is that these hypotheses must also allow for knowledge of the full profile of potential outcomes. Assuming that the treatment effect is additive and constant for all i, $y_{1i} = y_{0i} + \tau$, is a simple model that allows the imputation of all potential outcomes under the null hypothesis $H_0 : \tau = \tau_0$. The procedure to test this hypothesis is analogous to tests of the sharp null, except that first the observed outcomes are adjusted to subtract the hypothesized value $\tau = \tau_0$, and then the sharp null hypothesis is tested using the test statistic based on the adjusted outcomes, $W(\mathbf{T}, \mathbf{Y} - \mathbf{T}\tau_0) = W(\mathbf{T}, \mathbf{y}_0)$ (Rosenbaum, 2002b). Using this treatment effect model, we calculate confidence intervals by inverting hypothesis tests. That is, we construct a 95% confidence interval by testing the null hypothesis that $\tau = \tau_0$ for all possible values of τ_0, and keeping the hypotheses that we fail to reject at 5% level.

Naturally, if the constant treatment effect model is severely misspecified, the confidence intervals based on this model will not have correct coverage. For this reason, confidence intervals and equivalence tests based on this model ought to be interpreted with caution. We decided to adopt the constant treatment effect model despite its limitations because the number of observations in our application is limited and we did not have a strong theoretical expectation of heterogeneous treatment effects. We note, however, that it is possible to consider heterogeneous treatment effects in a Fisherian framework; indeed, the principle of constructing adjusted responses applies to very general models of treatment effects, even a model like $y_{1i} = y_{0i} + \tau_i$, where each observation has a different effect. Such a general model would of course have very limited practical use, because it would lead to an n-dimensional confidence set. One interpretation of the constant treatment effect model, offered by Rosenbaum (2010, §2.4.4), is that it is a simplification that sheds light on the n-dimensional effect $\theta = (\tau_1, \tau_2, \ldots, \tau_n)$, because we only exclude a scalar hypothesis τ_0 from a $1-\alpha$ confidence interval under the constant treatment effect model if and only if the n-dimensional hypothesis $\theta = (\tau_0, \tau_0, \ldots, \tau_0)$ is excluded from the n-dimensional confidence interval for θ. A

related way of exploring heterogeneous treatment effects in a Fisherian framework is by means of attributable effects (§2.5 Rosenbaum, 2010, 2001).

Finally, we note that our notation and analysis make the Stable Unit Treatment Value Assumption (SUTVA). When SUTVA holds, the outcome of every experimental unit is solely affected by the treatment received by that unit, regardless of the treatment status assigned to the rest of the units participating in the experiment (see,e.g., Rubin, 1990; Bowers et al., 2013). SUTVA fails when there is interference between units, as interference means that the potential outcome of each unit depends on the treatment status of other units. Fisherian tests of the sharp null hypothesis are still valid under interference; in this case, we can write the collection of i's potential outcomes as $y_i(\mathbf{t})$ for all possible $\mathbf{t}$, and the sharp null hypothesis as $H_0' : y_i(\mathbf{t}) = y_i(\mathbf{t}')$ for all $\mathbf{t}$, $\mathbf{t}'$ and for $i = 1, 2, \ldots, n$. Under H_0', it is still possible to impute all potential outcomes for all treatment assignments and hence derive the null distribution of test statistics, and a Fisherian randomization-based test still controls the probability of Type I error at the nominal level—although the interpretation of the test is more subtle, see Rosenbaum (2007). In contrast, our confidence intervals and tests of equivalence do rely on SUTVA because the treatment effect model we use to impute the missing potential outcomes assumes no interference. Thus, to simplify the exposition, we assume SUTVA throughout.

In our research design, SUTVA requires that a legislator who is reelection-ineligible behave in the same way regardless of how many other legislators in the chamber are reelection-ineligible. This would restrict scenarios where, for example, reelection-eligible legislators let reelection-ineligible legislators have a larger share of those resources that have a fixed budget (e.g. floor time) to help them take actions that will position them favorably in their quest for higher political office. However, given that reelection-ineligible legislators are not returning to the chamber, these agreements might be difficult to sustain in equilibrium (see Muthoo and Shepsle, 2010). Moreover, this kind of strategic coordination may be less likely to occur for outcomes that are not directly constrained by the actions of others (e.g. bill introductions).

Addressing Concerns About Power: Appropriate Test Statistics and Tests of Equivalence

In order to tackle the challenge of low power within the Fisherian framework, we employ both (i) different test statistics, (ii) and tests that invert the usual null hypothesis to control the probability

that the effect is incorrectly declared null. The latter tests are relevant in small-sample experiments like ours, where a failure to reject the null hypothesis of no effect can be driven by a lack of statistical power and thus cannot easily be interpreted as evidence that the effects are zero.

The Choice of Test Statistic.

Although the Fisherian framework allows us to derive the exact finite-sample distribution of any test statistic $W(\mathbf{T}, \mathbf{Y})$, some statistics will have more power to detect certain departures from the null hypothesis than others. For example, a common choice of test statistic is the difference in means between the treated and control outcomes, $W_{\texttt{DM}} = \frac{\sum_{i:T_i=1} Y_i}{n_1} - \frac{\sum_{i:T_i=0} Y_i}{n_0}$, which is appealing because of its straightforward average treatment effect interpretation in a Neyman (and also in a super-population) framework (Imbens and Rubin, 2015). $W_{\texttt{DM}}$ will be most powerful to detect departures from the sharp null hypothesis when the treatment induces a location shift, and less powerful when the treatment affects other features of the outcome distribution while location remains unchanged. Another limitation of this test statistic is that it is not robust to outliers.

Relative to the difference-in-means, test statistics based on ranks have the advantage that they are insensitive to outliers and may be more powerful to detect multiplicative effects (Imbens and Rubin, 2015, §5). Robustness to outliers is especially relevant in our study of Arkansas senators, because one of the outcomes we analyze below—abstentions—has an outlier observation that makes the difference-in-means statistic potentially misleading. For implementation, we follow Imbens and Rubin (2015) and employ the difference in the average ranks between treatment and control groups, defined as $W_{\texttt{DR}} = \frac{\sum_{i:T_i=1} R_i^y}{n_1} - \frac{\sum_{i:T_i=0} R_i^y}{n_0}$, where R_i^y is the rank of the observed response Y_i, for $i = 1, 2, \ldots n$. For alternative rank-based test statistics, see Rosenbaum (2002b, 2010).

Another strategy to address concerns about statistical power is to employ test statistics based on covariate adjustment (Rosenbaum, 2002a). The idea is simply to construct a test statistic based on the residuals from a fit of the outcome on one or more predetermined covariates, an approach that will be useful when such residuals are less dispersed than the unadjusted outcomes. The fit of the outcome on the covariates is purely algorithmic, not an assumed statistical model. In our analysis below, we employ a covariate adjusted version of $W_{\texttt{DM}}$, defined as $W_{\texttt{CDM}} = \frac{\sum_{i:T_i=1} \hat{e}_i}{n_1} - \frac{\sum_{i:T_i=0} \hat{e}_i}{n_0}$, where $\hat{e}_i$ is the residual obtained from a least-squares fit of the observed outcome on K predetermined covariates, $\hat{e}_i = Y_i - \mathbf{x}_i'\hat{\boldsymbol{\beta}}$, with $\mathbf{x}_i$ a $K \times 1$ vector of covariates, and $\hat{\boldsymbol{\beta}}$ a $K \times 1$ vector of least-squares

coefficients from a fit of $\mathbf{Y}$ on the matrix of covariates, $\mathbf{X} = [\mathbf{x}_1', \mathbf{x}_2', \ldots, \mathbf{x}_n']'$.

We also consider the Kolmogorov-Smirnov (KS) statistic, $W_{\texttt{KS}} = \sup_y |\hat{F}_1(y) - \hat{F}_0(y)|$, which measures the maximum absolute difference in the empirical cumulative distribution functions (CDF) of the treated and control outcomes—denoted respectively by $\hat{F}_1(\cdot)$ and $\hat{F}_0(\cdot)$. This test statistic has the advantage of capturing any difference in the distributions of treated and control groups, including not only differences in the means but also differences in other moments and in quantiles. We include it in our analysis to ensure that our failure to reject the sharp null hypothesis is not an artifact of using differences-in-means and rank-based statistics that may miss certain distributional effects.

Finally, when we analyze covariate balance, we also consider two omnibus test statistics that allow us to test the sharp null hypothesis for all covariates simultaneously: Hotelling's T^2 statistic and the maximum absolute difference in the t statistic across all covariates (Imbens and Rubin, 2015). Hotelling's T^2 is the squared Mahalanobis distance between the covariate averages in the treated and control groups, defined as $W_{\texttt{H}} = \mathbf{d}'(\hat{\boldsymbol{V}})^{-1}\mathbf{d}$, where $\bar{x}_{k1} - \bar{x}_{k0}$ is the treated-control difference in means for covariate k, $\mathbf{d} = (\bar{x}_{11} - \bar{x}_{10}, \bar{x}_{21} - \bar{x}_{20}, \ldots, \bar{x}_{K1} - \bar{x}_{K0})'$, and $\hat{\boldsymbol{V}}$ is the sample variance-covariance matrix of $\mathbf{d}$. The maximum t statistic is simply $W_{\texttt{Mt}} = \max(|t_1|, |t_2|, \ldots, |t_K|)$, where $t_k = \frac{\bar{x}_{k1} - \bar{x}_{k0}}{\sqrt{s_{k1}^2/n_1 + s_{k0}^2/n_0}}$ and s_{k1}^2 and s_{k0}^2 are, respectively, the sample variances of covariate k in the treated and control groups.

As we illustrate with our empirical application, this collection of alternative test statistics offers researchers the ability to build exact randomization-based tests that are also appropriately sensitive to the alternative hypotheses they consider most relevant. In our case, our concerns about low power make covariate adjustment an attractive choice; in addition, by considering rank-based statistics and statistics that capture any difference in distribution such as the KS statistic, we ensure that our tests of the sharp null hypothesis are robust to outliers and sensitive to non-additive treatment effects.

Testing the Null Hypothesis That The Treatment Effect is Nonzero

So-called tests of equivalence test the null hypothesis that two groups are different or a treatment effect is non-zero. Such tests are commonly used in medical studies to establish bioequivalence between generic and brand-name drugs (Berger et al., 1996). Letting μ_1 be the mean in the

treatment group and μ_0 the mean in the control group, tests of equivalence typically make the null hypothesis that the discrepancy between both means is larger than a positive number δ, and only reject the null hypothesis when there is sufficient evidence that the two groups are similar—that is, when there is evidence that both $\mu_1 - \mu_0 \geq -\delta$ and $\mu_1 - \mu_0 \leq \delta$.

We adapt this idea to a Fisherian framework, following related ideas in Rosenbaum (2010, §19). Our procedure involves simply adopting the constant treatment effect model, $y_{i1} = y_{i0} + \tau$, and inverting randomization-based tests of the null hypothesis about the parameter τ based on adjusted responses. We consider the null hypotheses $H_0^\delta : |y_{1i} - y_{0i}| > \delta$, for $i = 1, 2, \ldots, n$, and the two sub-hypotheses $H_0^{\delta,+} : y_{1i} - y_{0i} > \delta$ and $H_0^{\delta,-} : y_{1i} - y_{0i} < -\delta$. Given the constant treatment effect model, we have $H_0^\delta : |\tau| \geq \delta$, $H_0^{\delta,+} : \tau > \delta$ and $H_0^{\delta,-} : \tau < -\delta$. Our procedure tests both $H_0^{\delta,+}$ and $H_0^{\delta,-}$ using two one-sided randomization-based 5%-level tests, rejecting each hypothesis if the randomization-based p-value is at most 5%, and rejecting H_0^δ if both $H_0^{\delta,+}$ and $H_0^{\delta,-}$ are rejected. This procedure leads to a 5%-level test because the set of hypotheses $\left\{H_0^{\delta,-}, H_0^{\delta,+}\right\}$ is exclusive—i.e., it contains at most one true hypothesis (Rosenbaum, 2010).

Naturally, whether we reject H_0^δ depends on the value of δ. In our analysis below, we follow Rosenbaum (2010) and report the *maximum* value of δ for which H_0^δ fails to be rejected at 5% level—we refer to this value as $\delta^\star$. When we find $\delta^\star$, we can assert with 95% confidence that the shift τ in legislative output that occurs when a senator goes from being reelection-eligible to being reelection-eligible is at most $\delta^\star$, i.e. we can assert that $|\tau| \leq \delta^\star$. Thus, if $\delta^\star$ is small, we can rule out with 95% confidence that the effect of removing reelection incentives—under a constant treatment effect model—is large. The procedure to find $\delta^\star$ starts by testing H_0^δ for the maximum possible value for δ, $\delta = \infty$, and continues to test smaller values of δ until either $H_0^{\delta,-}$ or $H_0^{\delta,+}$ fails to rejected. The order is important: one must start with the largest value of δ and subsequently decrease it, to ensure that the sequence $\left\langle \left\{H_0^{\delta,-}, H_0^{\delta,+}\right\},\ \delta \in (0, \infty) \right\rangle$ is a sequentially exclusive partition of hypotheses—that is, a sequence where, for each value of δ, the set $\left\{H_0^{\delta,-}, H_0^{\delta,+}\right\}$ contains at most one true hypothesis when all the prior hypotheses are false. This ensures that the probability of rejecting at least one true hypothesis in the sequence of tests that leads to $\delta^\star$ is at most 0.05 (Rosenbaum, 2008).

5 Empirical Analysis

In this section, we first report covariate balance in our original sample, then present effects on the outcomes of interest, and finally report a bounds analysis to address sample attrition.

Covariate Balance in Original Experimental Sample

Under the random assignment of term length, the "treatment effect" is by construction zero for all senator-level predetermined characteristics. Thus, if we observed significant dissimilarities in predetermined covariates between our two samples, the validity of the randomization might be called into question.

We present the results from covariate balance tests based on our full sample of 64 senators. We test the sharp null hypothesis that there is no treatment effect on predetermined covariates using the randomization inference approach described above. We employ the difference-in-means between the reelection-ineligible and reelection-eligible groups as a test statistic, $W_{\texttt{DM}}$, and the two omnibus tests based $W_{\texttt{H}}$ and $W_{\texttt{Mt}}$ described above. We also report tests of equivalence based on the constant treatment effect model. We consider seven predetermined covariates, including the vote share obtained in the previous election, party, race, age and gender.

As shown in the first four columns of Table 1, we fail to reject the sharp null hypothesis for every single covariate (minimum p-value across all covariates is 0.18). We also fail to reject the null hypothesis when we conduct omnibus tests of covariate balance based on $W_{\texttt{H}}$ and $W_{\texttt{Mt}}$, as reported in the lower panel of the table. Some of these mean differences, however, are high. For example, 91% of reelection-ineligible senators are Democrat and 12% are Black, but the corresponding percentages in the reelection-eligible group are 75% and 6%. These large differences are partly driven by the combination of our low sample size and the relatively low frequency of Republican and Black senators. For example, 4 out of 32 senators are Black in the treatment group, while 2 out of 32 senators are Black in the control group. The difference is only 2 out of 6 senators and cannot be distinguished from chance in a binomial test, but because the denominator is only 32 in both cases, this difference is non-negligible in percentage points. A similar phenomenon occurs with the Democrat variable, as there are only 11 Republican senators in our sample of 64 senators—3 of these are in the treatment group and 9 in the control group.

That the large differences in the Democrat and Black variables are consistent with a valid randomization does not mean, of course, that these differences could not be affecting our conclusions. In a finite sample, covariate imbalances that occur due to chance will affect conclusions if these covariates are related in a systematic way to the outcomes of interest. We are not aware of particular empirical evidence that would suggest that this is a problem in our case, but we cannot rule it out. This illustrates another challenge of very small samples: large numerical imbalances in covariates can occur with high probability due to the low number of observations.

A related issue is that failure to reject the null hypothesis could be driven by a lack of power. For this reason, we also test the hypothesis that the reelection-ineligible and reelection-eligible groups are different, using the randomization-based tests of equivalence described above. We report the results in the last two columns of Table 1. We report $\delta^\star$, the maximum value of δ for which H_0^δ fails to be rejected at 5%, and $\delta^\star/\texttt{sd}$, where `sd` is the pooled standard deviation across the treated and control groups. The latter measure gives us a better idea of the size of $\delta^\star$. For example, given the constant treatment effect model $y_{i1} = y_{i0} + \tau$, the first row shows that we can assert with 95% confidence that the vote share of reelection-ineligible senators is shifted no more than ± 10.24 percentage points relative to reelection-eligible senators, an absolute difference that represents 0.53 pooled standard deviations, a moderately-sized effect. Consistent with our findings and discussion above, the values of $\delta^\star$ for the Democrat and Black variables are large, and represent 0.83 and 0.64 standard deviations, respectively. In sum, we fail to reject the sharp null hypothesis for every single covariate and in two omnibus tests, and we can assert, based on the constant treatment effect model, that the samples are not extremely different in terms of covariates—but we cannot rule out moderate or small differences.

Considering all the evidence, we conclude that there are no signs that the random assignment of senators to groups was faulty, but also that the small sample prevents us from being able to assert with confidence that the covariate differences are negligible. We report balance tests separately by session in Section A2 of the Supplemental Appendix. Our conclusions remain generally unchanged, although one of the two omnibus balance tests for the 2000/2002 cohort has an associated p-value below 5%.

Table 1: Covariate Balance Between Reelection-Ineligible and Reelection-Eligible Arkansas Senators, pooling 81st (1997-1998) and 86th (2007-2008) Legislative Sessions

	Means			**Test of no effect**	**Max δ failing to reject $H_0^\delta : \lvert\tau\rvert > \delta$**	
	Tr	Co	Difference	p-value	$\delta^\star$	$\delta^\star$/sd
Vote Share	89.27	86.8	2.46	0.6	10.24	0.53
Married	0.88	0.88	0	1	0.14	0.43
Male	0.94	0.81	0.12	0.23	0.25	0.76
Democrat	0.91	0.75	0.16	0.18	0.31	0.83
Black	0.12	0.06	0.06	0.67	0.19	0.64
Attorney	0.25	0.34	-0.09	0.55	0.29	0.62
Age	50.66	52.78	-2.12	0.45	6.64	0.6
Hotelling omnibus				0.3275		
Max abs. val. t-tstat				0.5394		
Sample Size	32	32				

Note: 'Tr' refers to treatment group of reelection-ineligible senators (assigned 4-year lot in 1992 or 2002), and 'Co' refers to control group of reelection-eligible senators (assigned 2-year lot in 1992 or 2002). The test of no effect reports randomization-based p-values corresponding to the sharp null hypothesis that the treatment of removing reelection incentives has no effect for any unit using the difference-in-means test statistic. Tests of the hypothesis H_0^δ reported in the last two columns are also randomization-based, assuming a constant treatment effect model as explained in the text and employing $W_{\texttt{DM}}$; `sd` is the pooled standard deviation across treated and control observations.

The Effects of Removing Reelection Incentives in the Arkansas Senate

We now study our main question of interest, whether reelection-ineligible state senators reduce their legislative output and participation relative to their reelection-eligible counterparts. To do so, we examine five dependent variables at the individual level: the number of bills introduced, the number of bills passed, the number of bills cosponsored, the abstention rate on roll-call votes, and the number of resolutions. For each senator, the abstention rate is measured as the number of votes in which the senator votes neither Yay nor Nay, divided by the total number of votes cast by the senator; for a given senator, the number of bills cosponsored is the number of bills introduced by other fellow senators that include the senator as one of the sponsors of the bill. We use the number of resolutions as an imperfect proxy for constituency service. While we would prefer a more conventional measure of constituency service (e.g. number of district staff, trips back to the district), they are not readily available. Instead, we use data on the resolutions that state senators file during the legislative session. As is typically the case with constituency service, these resolutions are devoid of ideological content; examples include recognizing the achievements of a citizen within their district or congratulating a local high school for its athletic accomplishments. We were unable to collect data on cosponsorship for the 1992 cohort, so all our analyses of bills cosponsored are from the 2007 (86th) session and include only senators from the 2002 cohort. Table

Table 2: Descriptive Statistics for Outcome Variables in Arkansas Senate, pooling 81st (1997-1998) and 86th (2007-2008) Legislative Sessions

	Min		Max		Mean		Median		SD		1st Quartile		3rd Quartile	
	Tr	Co	Tr	Co	Tr	Co	Tr	Co	Tr	Co	Tr	Co	Tr	Co
Abstentions	0	0	30.96	5.56	2.36	1.14	0.92	0.92	5.97	1.33	0.05	0.33	2.04	1.3
Resolutions	0	0	8.00	4.00	2.23	1.57	1.50	2.00	2.21	1.27	1.00	0.50	3.00	2.0
Bills introduced	5	5	45.00	55.00	22.88	21.48	25.00	19.00	11.50	11.75	10.50	14.00	29.75	26.0
Bills passed	3	1	35.00	27.00	14.58	13.35	15.50	12.00	8.33	7.25	7.25	7.00	21.25	17.0
Bills Cosponsored	25	21	66.00	61.00	36.69	40.85	32.00	38.00	11.62	13.18	28.00	29.00	41.00	52.0

Note: 'Tr' refers to treatment group of reelection-ineligible senators (assigned 4-year lot in 1992 or 2002), and 'Co' refers to control group of reelection-eligible senators (assigned 2-year lot in 1992 or 2002). Number of observations is 26 in treatment and 23 in control, except for bills cosponsored which includes 12 treated and 13 control observations.

2 presents descriptive statistics for all outcome variables.

We note that our outcome analysis includes fewer observations than the covariate balance analysis reported above due to sample attrition. By 1997 and 2007, the years when the 81st and 86st Legislative Sessions begin, 15 senators in our sample of 64 had left the chamber: 12 senators left between 1993 and 1997 and 3 senators left between 2002 and 2007. This leaves us with a remaining sample of 49 senators, whom we call "survivors." Any time attrition occurs in an experimental setting it raises concerns that the remaining subjects are not representative of the original experimental sample or population, which in turn can lead to invalid inferences (Gerber and Green, 2012). This would occur in our case if a senator's defeat or retirement before term limits become binding is affected by the term length assigned after reapportionment—i.e., by our treatment variable. In this section, we treat this attrition as random, but in the following subsection we consider the robustness of our results to deviations from this assumption.

Table 3 shows randomization-based results from tests of the sharp null hypothesis and randomization-based confidence intervals based on the constant treatment model. Sharp null p-values are calculated using four of the test statistics discussed above: $W_{\texttt{DM}}$ (difference in means), $W_{\texttt{CDM}}$ (covariate-adjusted difference-in-means), $W_{\texttt{DR}}$ (difference in average ranks), and $W_{\texttt{KS}}$ (the maximum distance between the empirical CDFs). In addition, we present three different confidence intervals employing the test statistics $W_{\texttt{DM}}$, $W_{\texttt{CDM}}$, and $W_{\texttt{DR}}$; as mentioned above, these confidence intervals are based on inversion of hypothesis tests $H_0 : \tau = \tau_0$ given the constant treatment effect model. As before, we pool observations across the two cohorts to maximize the number of observations. In Section A3 of the Supplemental Appendix, we present additional results based on the covariate adjusted KS statistic and the difference in inter-quartile ranges for the pooled sample, and also separate

Table 3: Test of Sharp Null Hypothesis and Confidence Intervals Based on Different Test Statistics for Outcome Variables in Arkansas Senate, pooling 81st (1997-1998) and 86th (2007-2008) Legislative Sessions

	P-value from Test of Sharp Null				**95% CI for Constant Effect**		
	$W_{\texttt{DM}}$	$W_{\texttt{CDM}}$	$W_{\texttt{DR}}$	$W_{\texttt{KS}}$	$W_{\texttt{DM}}$	$W_{\texttt{CDM}}$	$W_{\texttt{DR}}$
Abstentions	0.49	0.50	0.73	0.47	[-0.66,3.36]	[-0.68,3.36]	[-0.48,0.75]
Resolutions	0.22	0.20	0.47	0.31	[-0.38,1.71]	[-0.37,1.71]	[-0.99,1]
Bills introduced	0.65	0.62	0.48	0.34	[-4.87,7.63]	[-4.56,7.37]	[-3.99,8.99]
Bills passed	0.58	0.55	0.62	0.78	[-3,5.46]	[-2.93,5.33]	[-3,5.99]
Bills Cosponsored	0.41	0.41	0.44	0.55	[-14.26,6.16]	[-14.16,6.11]	[-15.99,5.99]

Note: P-values correspond to randomization-based test of the sharp null hypothesis that the treatment has no effect for any unit employing different test statistics defined in the text. The treatment is the removal of reelection incentives, and the tests are based on a comparison of reelection-ineligible senators (assigned 4-year lot in 1992 or 2002) and reelection-eligible senators (assigned 2-year lot in 1992 or 2002). Confidence intervals are calculated by inverting randomization-based hypothesis tests in a constant treatment effect model, employing different test statistics. The number of observations for bills cosponsored is 15 treated and 14 control, respectively.

analyses for the 1997 (81st) and 2007 (86th) sessions.

Table 3 shows a consistent pattern of null results. We fail to reject the sharp null hypothesis for every one of the five outcomes we consider across the four different test statistics (the p-values range from 0.20 to 0.78). Consistently, all 95% confidence intervals include zero. Based on this pattern, there is not enough empirical evidence to assert that the removal of reelection incentives affected legislative participation in the Arkansas Senate during the sessions we study.

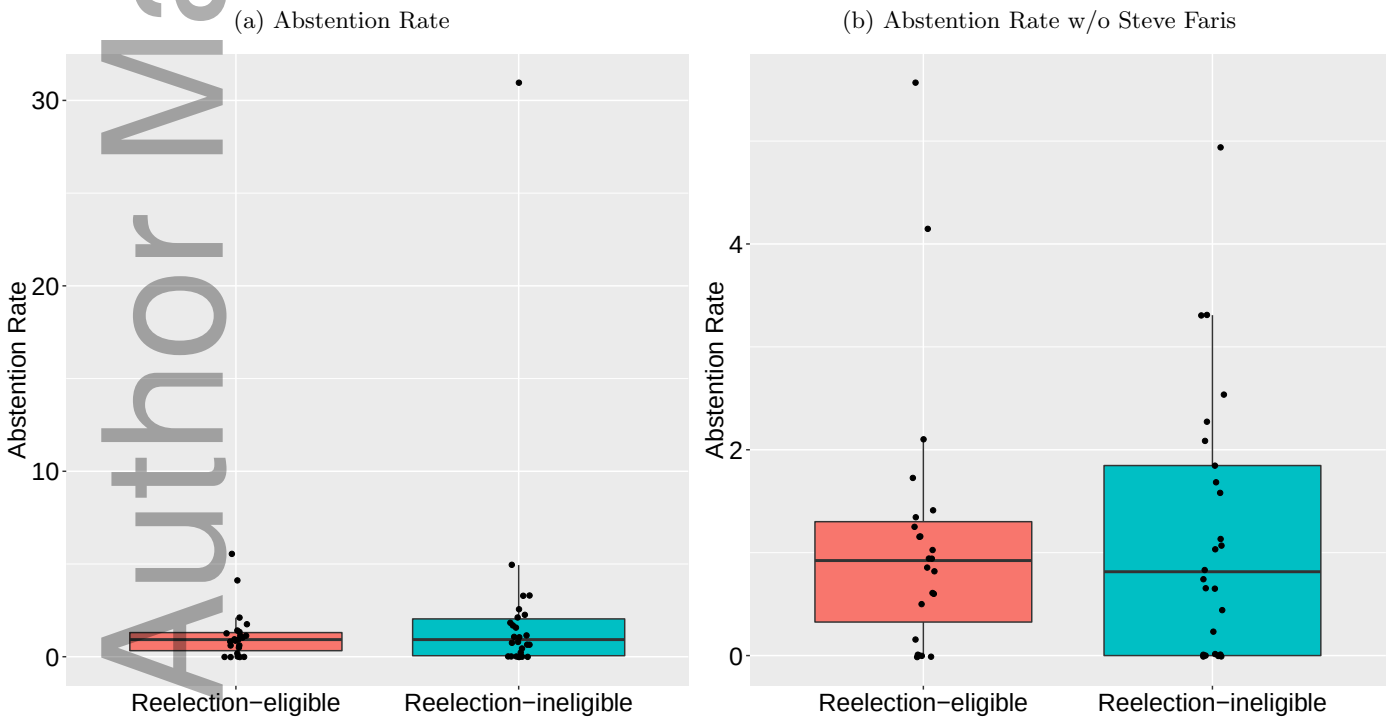
But there are further lessons to be learned from our results. The confidence interval for the abstention rate based on the difference in means, $W_{\texttt{DM}}$, is not symmetric around zero, ranging from a small negative effect of -0.66 percentage points to a larger positive effect of 3.36. This pattern is consistent with the averages reported in Table 2, where the abstention rate in the reelection-ineligible group (2.36 percent) is shown to be more than twice the rate in the reelection-eligible group. However, as illustrated in Figures 2(a) and 2(b), this mean difference is driven entirely by one senator in the treatment group, Steve Faris, who missed approximately 31 percent of roll-call votes during the 2007 session. As we discussed above, the difference in means is not robust to outliers, which causes the confidence intervals based on this statistic (and on its covariate-adjusted version, $W_{\texttt{CDM}}$) to be shifted to the right of zero. However, as shown in the last column of Table 3, when we employ the rank-based statistic, $W_{\texttt{DR}}$, the confidence interval becomes approximately symmetric about zero and its length is reduced by roughly 70%. This occurs because, unlike the difference in means, $W_{\texttt{DR}}$ is unaffected by the extreme value of the abstention rate exhibited by

Faris.

Moreover, covariate adjustment is shown to modestly reduce the length of confidence intervals in some cases. To compute the covariate-adjusted difference in means statistic, $W_{\texttt{CDM}}$, we employ the residuals from a least-squares fit of each outcome on previous vote share. For bills introduced, employing $W_{\texttt{CDM}}$ instead of $W_{\texttt{DM}}$ leads to a 4.5% reduction in confidence interval length. Similarly, for bills passed, covariate adjustment leads to confidence intervals that are 2.3% shorter. For the rest of the outcomes, the length of confidence intervals based on $W_{\texttt{CDM}}$ and $W_{\texttt{DM}}$ is very similar.

Finally, the fact that we reach the same conclusions when we employ $W_{\texttt{DM}}$, $W_{\texttt{CDM}}$, and $W_{\texttt{DR}}$ as when we employ the KS statistic, $W_{\texttt{KS}}$, suggests that our failure to reject the sharp null hypothesis using test statistics based on means and ranks is not because of these statistics' low power against non-shift alternatives, but rather because the entire outcome distributions of the reelection-ineligible and reelection-eligible groups are statistically indistinguishable.

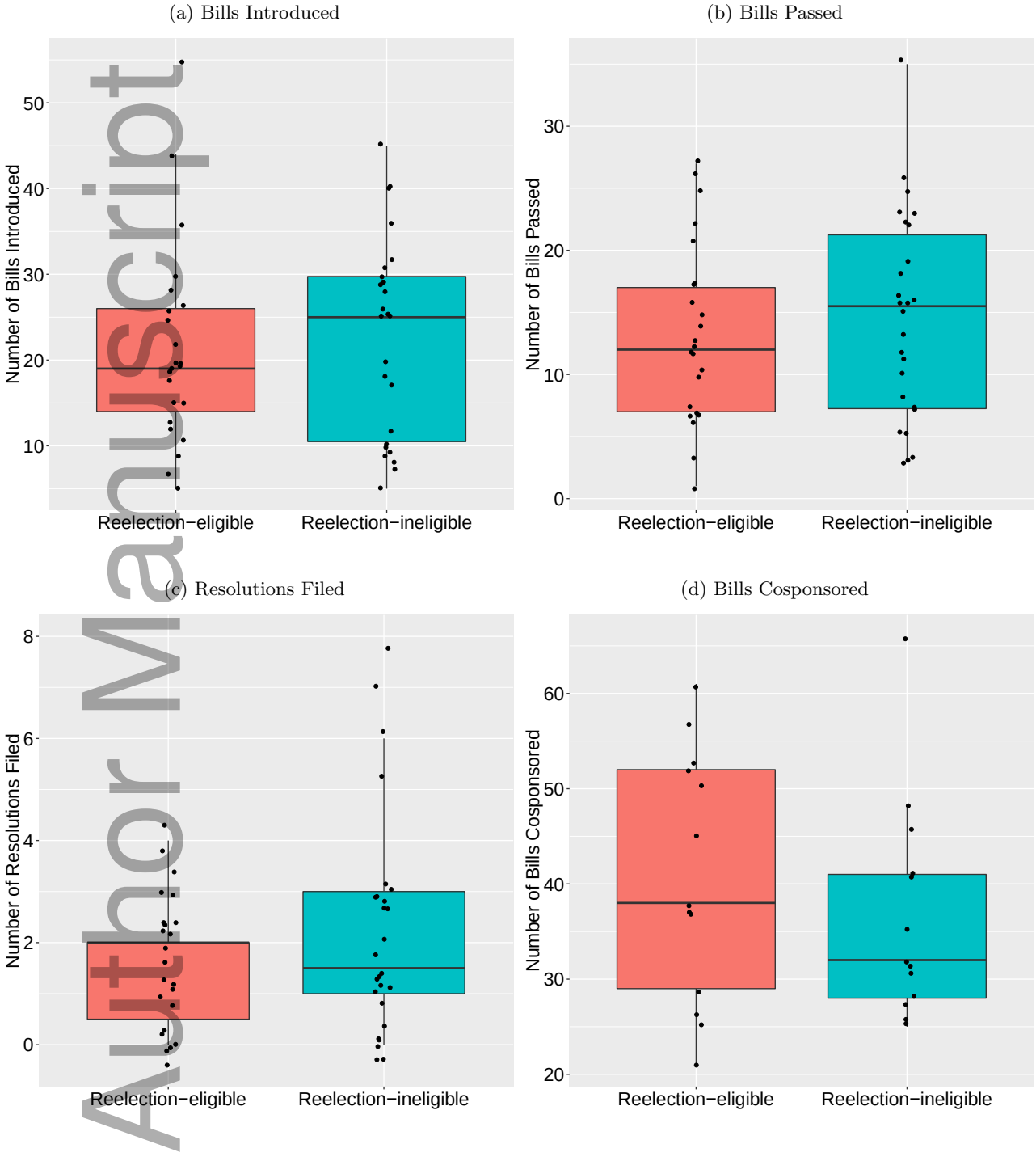
Figure 2: Removal of reelection effects on abstention rates in Arkansas Senate—81st (1997-1998) and 86th (2007-2008) Legislative Sessions



In sum, the randomization-based inferences presented in this section fail to provide evidence

that the absence of reelection incentives affects legislative output and participation. In particular, assuming random attrition, we do not find strong evidence that removing reelection incentives leads to participatory shirking. For some outcomes, mean participation is somewhat higher among reelection-ineligible senators. The box plots in Figures 3(a), 3(b), and 3(c) also show that, when looking at the entire distributions, there seems to be little evidence that the reelection-ineligible senators produce less legislative output in terms of bills introduced, bills passed, and resolutions. The pattern is somewhat different for bills cosponsored, as shown in Figure 3(d) and discussed below. However, we note that these results assume attrition occurs completely at random, and thus must be interpreted with caution.

Figure 3: Removal of reelection effects bill introduction, passage, and symbolic bills in Arkansas Senate—81st (1997-1998) and 86th (2007-2008) Legislative Sessions



Moreover, failing to reject the null hypothesis of no effect does not necessarily mean that we can be confident in asserting that the outcomes in the two groups are equivalent. This is true in

every application, and it is a more pressing concern in our case due to the low sample size, which affects our ability to detect true differences. We therefore present tests of equivalence, as we did for the covariates. As explained above, in these tests, our null hypothesis is that, in the constant treatment effect model $y_{1i} = y_{0i} + \tau$, τ is sufficiently large— $H_0^\delta : |\tau| > \delta$, for $\delta > 0$. In other words, our null hypothesis is that the legislative output of a senator under no reelection is equal to his/her output under reelection plus a large shift τ (which can be positive or negative), and we control the probability of incorrectly asserting that the treatment effect is small or negligible.

The first four columns of Table 4 report $\delta^\star$, the maximum value of τ that fails to reject the null hypothesis of nonequivalence, H_0^δ. We present results based on two different statistics, $W_{\texttt{DM}}$ and $W_{\texttt{DR}}$. As in Table 1, we report both $\delta^\star$ and $\delta^\star/\texttt{sd}$ for each test. Based on the results for $W_{\texttt{DM}}$, we can say with 95% confidence that $|\tau|$ is at most approximately 6 bill introductions, 4 bills passed, a 3 percent abstention rate, 1.5 resolutions and 13 bills cosponsored.

The results employing the rank-based statistic $W_{\texttt{DR}}$ lead to generally similar conclusions, with some important exceptions. For abstentions, $\delta^\star$ decreases dramatically from 3.08 to 0.65, illustrating once again that the conclusions based on the difference in means are heavily affected by the outlier observation in the treatment group. The second row of Table 4, which reports the results for abstention rates excluding Steve Faris, shows that the conclusions based on $W_{\texttt{DM}}$ and $W_{\texttt{DR}}$ are very similar when the outlier is excluded.

For resolutions, the rank-based value of $\delta^\star$ also decreases considerably relative to the value based on $W_{\texttt{DM}}$, from 1.5 to 1, but the results for bills introduced lead to a larger $\delta^\star$. In general, $\delta^\star$ ranges from a small effect of 0.15 standard deviations in the case of the abstention rate, to a large effect of 1.05 standard deviation for bills cosponsored, with most values between 0.5 and 0.7. (We note that the value of $\delta^\star/\texttt{sd}$ for abstentions using all observations is higher if we use the standard deviation of the control group instead. This is because the standard deviation in the treatment group is considerably higher due to the outlier observation. For example, for the results based on $W_{\texttt{DR}}$, $\delta^\star/\texttt{sd} = 0.15$ but it increases to 0.50 if we divide by the control standard deviation.)

As we discussed, we are particularly interested in whether we can rule out large *negative* effects associated with removing reelection incentives. As mentioned in the introduction, for several of our outcomes of interest, equivalence tests will be particularly informative about this one-sided, negative-effect hypothesis, because some test statistics have observed values that run counter to the

negative-effect hypothesis. For example, the observed mean differences show reelection-ineligible senators participating at equivalent or higher rates, not lower, for many of the outcomes we consider. Under these circumstances, tests of one-sided null hypotheses of non-equivalence will be most useful.

To explore this issue, we again employ the constant treatment effect model and test the hypothesis that a senator produces less legislative output if he is reelection-ineligible than if he is reelection-eligible—that is, we test the null hypothesis of shirking, $H_0^{\delta,S} : \tau < -\delta$, for $\delta > 0$. Note that $\tau = y_{i1} - y_{i0} < 0$ implies shirking only for the bill and resolution outcomes, but not for our abstention rates, since a *positive* τ for abstentions is a shirking effect. For this reason, for this outcome we define the null hypothesis of shirking as $H_0^{\delta,S} : \tau > \delta$, for $\delta > 0$.

The results, which we report in columns 5-8 of Table 4, show that tests of these one-sided hypotheses are considerably more informative than for the two-sided hypothesis $H_0^{\delta} : |\tau| > \delta$ for most of our outcomes, and allow us to rule out large negative effects of removing reelection incentives on legislative output and participation (under a constant treatment effect model). The results show that, in several cases, the $\delta^{\star}$ for the "shirking" hypothesis $H_0^{\delta,S}$ is considerably smaller than the $\delta^{\star}$ for the two-sided hypothesis of non-equivalence $H_0^{\delta} : |\tau| > \delta$. For example, the results for bills introduced based on $W_{\texttt{DR}}$ show that $\delta^{\star}$ decreases from 6.99 to 3.98 when we switch from H_0^{δ} to $H_0^{\delta,S}$, allowing us to assert that $\tau \geq -3.98$ and thus rule out negative effects that could not be ruled out based on a test of H_0^{δ}, which could only justify the assertion that $\tau \geq -6.99$. This is an improvement, but we note that, although four bill introductions is a relatively modest effect in terms of standard deviation, as a share of the average workload, it still constitutes a non-negligible effect.

A similar phenomenon occurs for resolutions and bills passed. For resolutions, the decrease in $\delta^{\star}$ that occurs when considering $H_0^{\delta,S}$ instead of H_0^{δ} is very large: from 1.5 to 0.25 when using $W_{\texttt{DM}}$, and from 1 to 0 when using $W_{\texttt{DR}}$. The results for resolutions reflect a phenomenon illustrated in Table 2 and Figure 3(c): senators in the reelection-ineligible group tend to introduce more resolutions than senators in the reelection-eligible group as measured by several aspects of their distributions (including means, and first and third quartiles). This leads to a rejection of almost all null hypotheses that assert that removing reelection incentives leads to a decrease in resolutions, i.e. null hypotheses that assert that $\tau < 0$.

The exception to this pattern is the number of bills cosponsored, reported in the last row.

Table 4: Tests of equivalence and negative effects (shirking), pooling 81st (1997-1998) and 86th (2007-2008) Legislative Sessions

	Max δ failing to reject $H_0^{\delta}: \lvert\tau\rvert > \delta$				**Max δ failing to reject $H_0^{\delta,S}: \tau < -\delta$**			
	$W_{\texttt{DM}}$		$W_{\texttt{DR}}$		$W_{\texttt{DM}}$		$W_{\texttt{DR}}$	
	$\delta^\star$	$\delta^\star$/sd	$\delta^\star$	$\delta^\star$/sd	$\delta^\star$	$\delta^\star$/sd	$\delta^\star$	$\delta^\star$/sd
Abstentions	3.08	0.69	0.65	0.15	3.1	0.7	0.65	0.15
Abstentions wo SF	0.69	0.53	0.54	0.41	0.7	0.53	0.54	0.41
Resolutions	1.5	0.82	1	0.54	0.25	0.13	0	0
Bills introduced	5.97	0.51	6.99	0.6	4.41	0.38	3.98	0.34
Bills passed	4.34	0.55	4	0.51	2.74	0.35	3	0.38
Bills Cosponsored	12.59	1.01	13.08	1.05	12.65	1.02	13	1.05
Sample Size	26	23						

Note: Tests of the hypotheses H_0^{δ} and $H_0^{\delta,S}$ are performed using randomization inference, assuming a constant treatment effect model and employing different test statistics as explained in the text; sd is the pooled standard deviation across treated and control observations. The treatment is the removal of reelection incentives, and the tests are based on a comparison of reelection-ineligible senators (assigned 4-year lot in 1992 or 2002) and reelection-eligible senators (assigned 2-year lot in 1992 or 2002). Calculations for abstention rates excluding Steve Faris (in the reelection-ineligible group) use a total sample size of 48. In the last four columns, the results for Abstentions and Abstentions without Steve Faris correspond to tests of $H_0^{\delta,S}: \tau \geq \delta$. The number of observations for bills cosponsored is 15 treated and 14 control, respectively.

Because the reelection-ineligible group has a lower average of cosponsored bills than the reelection-eligible group (see last row in Table 2), the $\delta^\star$ associated with the two-sided H_0^{δ} is roughly the same as the $\delta^\star$ associated with the negative-effect one-sided $H_0^{\delta,S}$. This outcome thus illustrates that our ability to rule out large negative effects is tightly connected to the distribution of the observed treated and control outcomes. In particular, when responses tend to be higher in the treated than in the control group, testing the non-equivalence hypothesis that the treatment decreases responses in a constant treatment effect model will lead to more informative (i.e., smaller) values of $\delta^\star$. But this will not occur when treated responses tend to be lower than control responses, as occurs for bills cosponsored.

Bounds to Assess Robustness to Sample Attrition

The previous sections assumed that missing observations were missing completely at random. In this section, we explore whether some of our conclusions survive patterns of retirement or defeat that may be correlated with the initial assignment of term length. First, we note that attrition rates are comparable across treatment and control groups. Our initial sample size is 32 senators in each group, and after attrition there are 26 senators in the reelection-ineligible group and 23 senators

in the reelection-eligible group. The difference in the non-missingness rates by group (23/32 and 26/32) are statistically indistinguishable (p-value is 0.3840), and the null hypothesis that the true probability of success is equal to 0.5 in 49 trials of a Bernoulli experiment cannot be rejected with 23 successes (p-value 0.7754). This balance in attrition rates is also seen when we consider each legislative session individually. In the 1990s cohort, 8 senators assigned 2-year terms and 4 senators assigned 4-year terms drop out of the sample before 1997, and in the 2000s cohort 1 senator assigned a 2-year term and 2 senators assigned 4-year terms drop out of the sample before 2007.

We also note that, in addition to the initial randomization, which allows us to ensure comparability at baseline, a crucial aspect of our design is that both groups of senators have survived the same number of elections—one—when the outcomes are observed. As a result, the attrition that results from electoral defeat in the first reelection is likely to affect both groups equally, and the composition of both groups in terms of "departors" (senators who drop out before term limits are binding and whose outcomes we fail to observe) and "survivors" (senators whose outcomes we observe) is likely to be similar at the moment when outcomes are measured. Moreover, this composition is equal (on average) at baseline due to the initial randomization.

Nonetheless, and despite losing roughly the same number of senators in each group, there could still be differences in the *type* of senators who drop out. The distribution of predetermined covariates in the subsample of survivors suggests that senators in the reelection-ineligible group who survived until their last term may be different from senators in the reelection-eligible group who survived until their penultimate term. As shown in Table 5, although we fail to reject the sharp null hypothesis for every single covariate and the omnibus hypothesis of balance at 5%, the treated-control differences in predetermined covariates in the survivor sample are larger than in the full sample (reported in Table 1) and values of $\delta^\star$ also increase. These results suggest that ignoring attrition may lead to incorrect conclusions.

To address this issue, we estimate upper and lower bounds on the average treatment effect following Manski (2003). Our focus on this parameter represents a shift away from the Fisherian framework, where null hypotheses about the average treatment effect are typically not studied because they are not sharp—i.e., they do not allow the imputation of all missing potential outcomes that are needed to calculate the exact distribution of test statistics under the null. A focus on the average effect, however, is consistent with a Neyman framework that treats potential outcomes

Table 5: Covariate Balance Between Reelection-Ineligible and Reelection-Eligible Arkansas Senators, pooling 81st (1997-1998) and 86th (2007-2008) Legislative Sessions—Only survivors

	Means			**Test of no effect**	**Max δ failing to reject $H_0^\delta : \lvert\tau\rvert > \delta$**	
	Tr	Co	Difference	p-value	$\delta^\star$	$\delta^\star/\texttt{sd}$
Vote Share	90.79	83.66	7.13	0.18	16.99	0.87
Married	0.85	0.91	-0.07	0.66	0.22	0.67
Male	0.96	0.78	0.18	0.09	0.33	1.03
Democrat	0.96	0.78	0.18	0.09	0.33	1.03
Black	0.15	0.04	0.11	0.36	0.27	0.87
Attorney	0.27	0.39	-0.12	0.48	0.27	0.57
Age	50.81	50.74	0.07	0.99	5.56	0.49
Hotelling omnibus				0.1003		
Max abs. val. t-tstat				0.3746		
Sample Size	26	23				

Note: 'Tr' refers to treatment group of surviving reelection-ineligible senators (assigned 4-year lot in 1992 or 2002), and 'Co' refers to control group of surviving reelection-eligible senators (assigned 2-year lot in 1992 or 2002). The test of no effect reports randomization-based p-values corresponding to the sharp null hypothesis that the treatment of removing reelection incentives has no effect for any unit using the difference-in-means test statistic. Tests of the hypothesis H_0^δ reported in the last two columns are also randomization-based, assuming a constant treatment effect model as explained in the text and employing $W_{\texttt{DM}}$; `sd` is the pooled standard deviation across treated and control observations.

as fixed and computes expectations in the finite population (Imbens and Rubin, 2015). We thus understand the analysis in this section as computing bounds on the finite population average treatment effect given that the finite population has two subgroups—survivors and departors—where the share of observations in each subgroup is known but the outcomes in the departor group are not.

To calculate the upper bound on the average treatment effect, we set the outcome values of those senators initially assigned to the treatment group who do not survive until the 1997 or 2007 sessions equal to the 75th percentile of the outcome values in our pooled survivor sample, and the missing outcome values of those senators initially assigned to the control group equal to the 25th percentile of the observed survivor outcomes. To calculate the lower bound on the average treatment effect, we do the opposite, setting missing outcomes in the treatment group at the 25th percentile value of the observed outcomes, and missing outcomes in the control group at the 75th percentile value. In other words, our lower bound assumes that all missing outcomes in the reelection-ineligible group, if observed, would have been low while all the missing outcomes in the reelection-eligible group would have been high. Analogously, our upper bound assumes that all missing outcomes in the treatment group, if observed, would have been high while all the missing outcomes in the control

group would have been low.

Table 6 shows the bounds on the average treatment effect—the average potential outcome under no possibility of reelection minus the average potential outcome under the possibility of one more reelection—for each outcome. Since our intention is to assess the robustness of our conclusion that removing reelection incentives does not lead to large negative effects on legislative participation, we focus on the lower bound for bill and resolution outcomes and on the upper bound for abstention rates. The results in Table 6 show that for bill introductions, bill passage, resolutions and abstention rates, even a severely endogenous pattern of attrition would result in a relatively small shirking or negative effect as measured by the average treatment effect. For example, for the number of bills introduced, assuming that all missing reelection-ineligible senators would have introduced just 12 bills (25th percentile) while all missing reelection-eligible senators would have introduced 29 bills (75th percentile) would result in a lower bound for the average effect of just -2.75 bills, showing that even under severely endogenous attrition we could rule out large average effects on this outcome.

A similar pattern is observed for bills passed, resolutions and abstention rates. The lower bound on the last-term effects on bill passage is just -1.78 bills, and this is assuming that missing reelection-ineligible senators would have passed just 7 bills while missing reelection-eligible senators would have passed 19. The lower bound on resolutions is positive at 0.03, ruling out negative effects. And the upper bound on last-term effects for abstention rates excluding Steve Faris is 0.43%, less than half of a percentage point, assuming that the missing abstention rates among reelection-ineligible senators would have been 1.68% while the abstention rate among missing reelection-eligible senators would have been 0.16%, about ten times smaller. (As shown, including senator Faris increases this bound to 1.37.)

The exception, once again, is bills cosponsored. Consistent with our prior results, for this outcome, we find that the lower bound of the average treatment effect is -5.75 and the upper bound bound is -1.81, indicating that for the 2000/2002 cohort, even under our most optimistic assumptions about attrition, removing reelection incentives leads to a decrease in the average number of bills cosponsored. We note that these results are consistent with the statistically insignificant results for cosponsorship in Table 3, because our bounds are point estimates and we are not reporting randomization/sampling uncertainty.

Table 6: 75th and 25th Percentile Bounds on Average Treatment Effect of Reelection Ineligibility in Arkansas Senate, pooling 81st (1997-1998) and 86th (2007-2008) Legislative Sessions

	ATE Lower Bound	ATE Upper Bound
Abstentions	0.65	1.37
Abstentions wo SF	−0.27	0.43
Resolutions	0.03	0.97
Bills introduced	−2.75	5.22
Bills passed	−1.78	3.84
Bills Cosponsored	−5.75	−1.81

Note: Columns labeled 'ATE Lower Bound' and 'ATE Upper Bound' report, respectively, the estimated lower and upper bounds of the average treatment effect—the average potential outcome under the reelection ineligibility (treated) condition minus the average potential outcome under the reelection eligibility (control) condition. Upper bounds sets missing treated outcomes to 75th percentile of observed treated outcome and missing control outcomes to 25th percentile of observed control outcomes. Lower bound is analogous, using 25th percentile for missing treated and 75th percentile for missing control. Calculations for abstention rates excluding Steve Faris (in treatment group) use a total sample size of 63; the number of observations for bills cosponsored is 15 treated and 14 control, respectively.

6 Conclusion

We examine how removing reelection incentives affects legislative behavior in the Arkansas Senate, specifically the extent to which reelection-ineligible legislators produce less legislation and abstain at higher rates than their reelection-eligible counterparts. Our research design is based on two natural experiments in the Arkansas Senate that randomly assign term length immediately after reapportionment, and induce a change in the maximum number of terms that senators are allowed to serve. Since the inability to run for reelection is a central component of term limits policies, our study contributes to the broader literature on the effects of term limits in state legislatures. However, because our research design is based on a manipulation of the ability to run for reelection in a legislature where all senators are term limited, our study cannot be informative about the overall effect of adopting term limits, such as the effects that term limits are likely to have on the composition of the candidate pool.

Most of our results are based on a Fisherian framework where potential outcomes are seen as fixed and hypothesis tests are based on the randomization distribution of the treatment assignment, leading to the exact null distribution of test statistics and hence hypothesis tests that are finite-sample exact. The ability of the Fisherian framework to produce tests of correct size is appealing for our study, because our small sample size raises questions about the validity of large-sample

approximations.

Our low sample size also raises concerns about low statistical power, which is particularly problematic in our case because we find null results and we intend to use those null results as the basis of informative scientific claims. To address power concerns, we employ two strategies. First, we use various test statistics that are well equipped to detect different kinds of departures from the sharp null hypothesis of no effect. Second, we employ a constant treatment effect model embedded in a Fisherian framework to test the null hypothesis that the effect of treatment is higher than a certain threshold. These so-called equivalence tests invert the usual null hypothesis and allow us to control the probability of claiming that the treatment effect is null when in fact it is not. Since our sample is small, these tests lead us to rule out that the treatment effect is very large (larger than one pooled standard deviation) but not that the effect is moderate or small. For this reason, we test the one-sided null hypothesis that the treatment effect is negative, i.e. reelection ineligibility leads to less legislative output and higher abstentions. This hypothesis of last-term shirking has been at the center of prior studies.

We find that one-sided tests are more informative, and for most outcomes allow us to rule out large and moderate effects. As we discuss, this occurs because the theoretical expectation in the shirking literature is that the effects are negative, and for many outcomes the observed values of the test statistics run counter to those expectations. Under these two conditions, one-sided tests of equivalence allow researchers to rule out many effects in the theoretically expected direction, leading to informative conclusions even with a small sample. Our analysis of bills cosponsored also shows that when these conditions do not hold, the conclusions we can draw from a small sample are considerably more limited.

Finally, we also address concerns about non-random attrition, evaluating how the observed average differences between the reelection-eligible and reelection-ineligible groups would change if the outcomes of senators who retire or are defeated before they reach their term limits are systematically lower or higher in one group relative to the other. This analysis shows that, for most outcomes, the average treatment effect would not be large and negative even under severely systematic differences in the types of senators who survive in each group. The exception is bills cosponsored, an outcome for which even the most optimistic assumptions about attrition lead to a negative average effect.

Our results are necessarily limited in scope because they cover only one state, and additional empirical work is needed to ensure that our conclusions hold for non-participatory outcomes and are generalizable beyond Arkansas. Indeed, the Fisherian framework we adopt treats the Arkansas senators as the universe of analysis, with no assumption of random sampling. The external validity of our results is therefore necessarily limited, and we cannot conclude based on these results alone that the negative effects of removing reelection incentives on legislative participation would be small in all cases. Instead, we can make the more limited claim that we have found one political environment where several legislative participation outcomes (though not all) seem not to be negatively affected to a large degree by the removal of reelection incentives when compared to the possibility of being reelected one more time. As we said, from this it cannot be concluded that there are no negative consequences of adopting term limits, because our design fails to capture several important phenomena, such as potential changes to the overall composition of the legislature and the reduced incentives for long-term investments in policy expertise, all of which could have harmful consequences for the quality of representation. Moreover, we cannot rule out small negative effects.

Methodologically, we believe our study illustrates the usefulness of several statistical tools in the empirical analysis of social science applications. In particular, randomization inference tools avoid the need to use large-sample approximations that can be unreliable in small samples, and tests of equivalence provide a more informative analysis of null effects, providing researchers with a way to quantify the equivalence between groups or the extent to which treatment effects are negligible. Moreover, a partial identification analysis is helpful to address robustness to sample attrition in a fully nonparametric way. We believe a more frequent use of these tools in social science applications would be beneficial. In particular, if equivalence tests were used to complement the analysis of social science studies that find null effects, these studies would become richer and more informative, and some of the well-documented publication bias against them (e.g., Franco et al., 2014) might be avoided.

References

Berger, R. L., J. C. Hsu, et al. (1996). Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statistical Science 11*(4), 283–319.

Bowers, J., M. M. Fredrickson, and C. Panagopoulos (2013). Reasoning about interference between units: a general framework. *Political Analysis 21*(1), 97–124.

Carey, J., R. Niemi, L. Powell, and G. Moncrief (2006). The effects of term limits on state legislatures: A new survey of the 50 states. *Legislative Studies Quarterly 31*(1), 105–134.

Carey, J. M., R. G. Niemi, and L. W. Powell (1998). The effects of term limits on state legislatures. *Legislative Studies Quarterly 23*(2), 271–300.

Cattaneo, M. D., B. Frandsen, and R. Titiunik (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the u.s. senate. *Journal of Causal Inference 3*(1), 1–24.

CQ Press (2016). *Voting and Elections Collection*. Sage Publishing, Online.

Fearon, J. D. (1999). Electoral accountability and the control of politicians: Selecting good types versus sanctioning poor performance. In B. Manin, A. Przeworski, and S. Stokes (Eds.), *Democracy, Accountability, and Representation*. New York: Cambridge University Press.

Fisher, R. A. (1935). *Design of Experiments*. New York: Hafner.

Franco, A., N. Malhotra, and G. Simonovits (2014). Publication bias in the social sciences: Unlocking the file drawer. *Science 345*(6203), 1502–1505.

Gerber, A. and D. Green (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York: W.W. Norton and Company Inc.

Gerber, B. J. and P. Teske (2000). Regulatory policymaking in the american states: A review of theories and evidence. *Political Research Quarterly 53*(4), 849–886.

Glazer, A. and M. Wattenberg (1996). How will term limits affect legislative work? In B. Grofman (Ed.), *Legislative Term Limits: Public Choice Perspectives*, Boston, MA. Kluwer Academic Publishers.

Ho, D. E. and K. Imai (2006). Randomization inference with natural experiments. *Journal of the American Statistical Association 101*(475), 888–900.

Imbens, G. W. and P. R. Rosenbaum (2005). Robust, accurate confidence intervals with a weak instrument: Quarter of birth and education. *Journal of the Royal Statistical Society: Series A (Statistics in Society 168*, 109–126.

Imbens, G. W. and D. B. Rubin (2015). *Causal Inference in Statistics, Social, and Biomedical Sciences*. Cambridge University Press.

Mansbridge, J. (2009). A selection model of representation. *Journal of Political Philosophy 17*(4), 369–398.

Manski, C. (2003). *Partial Identification of Probability Distributions*. New York: Springer-Verlag.

Manski, C. F. (2007). *Identification for prediction and decision*. Harvard University Press.

Muthoo, A. and K. A. Shepsle (2010). Information, institutions and constitutional arrangements. *Public choice 144*(1), 1–36.

National Conference of State Legislatures (2009). Full and part-time legislatures. Accessed on November 15, 2013.

Nicholson-Crotty, S. (2009). The politics of diffusion: Public policy in the american states. *The Journal of Politics 71*(1), 192–205.

Powell, L., R. Niemi, and M. Smith (2007). Constituent attention and interest representation. In K. Kurtz, B. Cain, and R. Niemi. (Eds.), *Institutional Change in American Politics: The Case of Term Limits*, Ann Arbor, MI. University of Michigan Press.

Rosenbaum, P. R. (2001). Effects attributable to treatment: Inference in experiments and observational studies with a discrete pivot. *Biometrika 88*(1), 219–231.

Rosenbaum, P. R. (2002a). Covariance adjustment in randomized experiments and observational studies. *Statistical Science 17*(3), 286–327.

Rosenbaum, P. R. (2002b). *Observational Studies* (2nd ed.). New York: Springer-Verlag.

Rosenbaum, P. R. (2007). Interference between units in randomized experiments. *Journal of the American Statistical Association 102*(477), 191–200.

Rosenbaum, P. R. (2008). Testing hypotheses in order. *Biometrika 95*(1), 248–252.

Rosenbaum, P. R. (2010). *Design of Observational Studies*. New York: Springer-Verlag.

Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science 5*(4), 472–480.

Sarbaugh-Thompson, M. (2010). Measuring 'term limitedness' in us multi-state research. *State Politics & Policy Quarterly 10*(2), 199–217.

Sarbaugh-Thompson, M., L. Thompson, C. Elder, J. Strate, and R. Elling (2004). *The Political and Institutional Effects of Term Limits*. New York: Palgrave MacMillan.

Sekhon, J. and R. Titiunik (2012). When natural experiments are neither natural nor experiments. *American Political Science Review 106*(1), 35–57.

Shipan, C. R. and C. Volden (2006). Bottom-up federalism: the diffusion of antismoking policies from u.s. cities to states. *American journal of political science 50*(4), 825–843.

Titiunik, R. (2016). Drawing your senator from a jar: Term length and legislative behavior. *Political Science Research and Methods 4*(2), 293–316.

Will, G. (1992). *Restoration: Congress, Term Limits and the Recovery of Deliberative Democracy*. New York: The Free Press.