ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2018)
181, *Part* 2, *pp.* 351–378

Legislative behaviour absent re-election incentives: findings from a natural experiment in the Arkansas Senate

Rocío Titiunik

University of Michigan, Ann Arbor, USA

and Andrew Feher

Covered California, Sacramento, USA

[Received July 2015. Final revision March 2017]

Summary. We analyse the effect of removing re-election incentives on the individual legislative participation of state lawmakers by 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 cannot 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: Bounds; Fisherian inference; Legislative behavior; Term limits; Tests of equivalence

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 changes in policy in state government in decades. Between 1992 and 1996, 17 more states followed suit. At April 2015, 15 states jointly housing 37% of the total US population still limit state legislators' length of service. The consequences of these re-election 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 re-election incentives leads legislators, for example, to decrease their effort or to adopt 'out-of-step' ideological positions.

We study the effect of removing re-election 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 are 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

Address for correspondence: Rocío Titiunik, Department of Political Science, University of Michigan, 5700 Haven Hall, 505 South State Street, Ann Arbor, MI 48109, USA.
E-mail: titiunik@umich.edu

0964–1998/18/181351

still being adopted and modified in many states, their effects on legislative behaviour 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 influence 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' length of term 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 who are randomly assigned 4-year terms in the first election after reapportionment see one fewer session in office than those who are randomly assigned 2-year terms, which enables us to examine two legislative sessions in 1997 and 2007 where the group of legislators who were assigned 4-year lots is ineligible for re-election whereas the group that is assigned 2-year lots is still eligible for one last re-election. 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 re-election incentives on several measures of individual legislative participation and output.

Although the effects that we study are informative about a crucial aspect of term limits—the removal of re-election 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 re-election 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 (for example, term limits may discourage certain types of candidate from entering the race), and the effect that the removal of re-election incentives would cause on the behaviour of individual legislators. Our natural experiment stands in contrast with this ideal experiment, as it manipulates how many terms a senator can serve, in a given legislature, before becoming ineligible to run for re-election. Our treatment group comprises senators who are serving their last term and are not eligible for re-election when the term that they are serving expires; our control group comprises senators who are eligible for one last re-election at the end of the term that they are serving. The comparison of these groups captures the effect of removing re-election incentives relative to a condition where re-election incentives are in place for one additional electoral cycle. Our natural experimental design is thus informative about re-election incentive effects, but not about legislature level effects that would require a design that is akin to the ideal experiment.

Our data have 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 35 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 that is 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 enables 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 (2002a).

Although a Fisherian approach leads to tests of correct size, it does not solve the problem of low statistical power that is 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 that are tailored to detecting different types of departure 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 cannot reject the null hypothesis of no last-term effects and we are interested in asserting that the absence of re-election incentives does not alter individual legislative output, we test the null hypothesis that the outcomes of re-election ineligible lawmakers *differ* from those of re-election 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 and Hsu, 1996). Thus, when controling the type I error rate, we control the probability of declaring that removing re-election incentives has no effect when in fact it does. Following related ideas that were 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 we use it to calculate, for every outcome, the minimum discrepancy between the re-election ineligible and re-election eligible group that leads to a rejection of the null hypothesis that the effect of removing re-election is non-zero. As we show, these tests of equivalence allow us to assert with 95% confidence that the removal of re-election incentives induced by term limits does not have very large negative effects on most measures of individual legislative output and participation; however, we cannot rule out moderate-to-small negative effects.

Our analysis reveals that, in low power settings, equivalence tests are most informative when two conditions hold: theoretical expectations are one sided and 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 re-election incentives may induce legislators to exert low effort or to 'shirk', leading to the expectation that removing re-election will lead to lower legislative output. Moreover, for all except one of the individual legislative outcomes that we analyse, the average output in the re-election ineligible group is higher than in the re-election 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 to assert that the effects are null or small is reduced. Our analysis of the number of bills that were cosponsored by each legislator illustrates this phenomenon: since re-election ineligible senators cosponsor on average fewer bills than their re-election 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 by using the partial identification framework that was developed by Manski (2003, 2007). We shift our focus to the average treatment effect of removing re-election incentives and estimate bounds on this effect under the assumption that those lawmakers whose outcomes we do not 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 re-election 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 behaviour 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 on-line supplemental appendix.

The data that are analysed in the paper and the programs that were used to analyse them can be obtained from

`http://wileyonlinelibrary.com/journal/rss-datasets`

2. Theoretical expectations: legislative behaviour absent re-election incentives

Under an accountability model of representation, voters incorporate politicians' past actions in their voting decisions and elections are a mechanism to sanction representatives' behaviour. In turn, the threat of punishment induces re-election-seeking politicians to behave in accordance with constituents' preferences and expectations. Under this model, the logical consequence of removing the possibility of re-election is to induce undesirable legislative behaviour 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 re-election should result in systematic changes in legislative behaviour (Fearon (1999), page 63).

Moreover, there may be mechanisms by which the removal of re-election 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 harbour 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 participate actively in legislative activities or curb some of that legislative participation to invest attention in surveying their future employment options. Given the time commitments that are associated with casework and with building the coalitions that are needed to navigate bills successfully through the legislature, we might expect that those who are in their last term, and thus more likely to be in search of their next job, reduce the effort that they expend on constituency service and policy making.

These two non-exclusive scenarios—shirking induced by the removal of re-election incentives and shirking induced by the opportunity costs of securing future employment—imply that re-election ineligible legislators will have less time or incentives to meet staff to help to draft legislation, to learn about the kinds of bills that 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, re-election 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 re-election eligible members.

Last-term effects may also depend on the degree of the legislature's professionalism. Facing low pay, limited staff resources, poor prospects of advancement and the absence of a re-election incentive, members serving in one of the 24 states with low professionalization legislatures have few incentives to participate actively. By contrast, in professional legislatures, not only do staff subsidize the cost of policy making, but also 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 'red–white–blue' trifurcation, Arkansas is considered a 'white', or hybrid, legislature on the basis of 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 60 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 that was induced by more lenient term limit policies did not represent such a drastic change (Sarbaugh-Thompson (2010), page 202).

However, countervailing forces may attenuate the potential effects of removing re-election incentives. First, some term limits advocates expected that, by introducing term limits, lawmakers would be less preoccupied with re-election-centred activities and would expend more effort on policy making and related legislative activities (Will, 1992; Glazer and Wattenberg, 1996). If this argument is true, the extra available time that is enjoyed by re-election ineligible legislators should more than compensate for the incentives to expend less effort. Second, most senators in our sample are elected or re-elected in very lopsided elections (the average vote share in our sample is 88%), suggesting that re-election pressures will be lower than in other more competitive settings. Third, the need to secure future employment is probably 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 that the future higher office election would act as a strong incentive even when re-election 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 US House seats (Steve Bryles, Gene

Jeffress, Jay Bradford and Vic Snyder), four ran for US Senate seats (Gilbert Baker, Kim Hendren, Jim Holt and Lu Hardin) and three ran for Governor (Shane Broadway, Jim Holt and Mike Beebe)—see CQ Press (2016).

These countervailing forces may lead to the expectation that removing re-election incentives in the Arkansas Senate will have null or positive effects on legislative output. For this reason, all the initial statistical tests that we present are two sided. At the same time, we are particularly interested in the hypothesis that the removal of re-election incentives has negative implications for the effort and legislative output of individual members, both because we wish to rule out—or to show that we cannot rule out—some of the potential negative normative implications of removing re-election incentives, and because previous non-experimental studies of term limits in the US's state legislatures 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) found that term-limited state legislators devote less time to securing rewards for their district and helping constituents to deal with government. In addition, Sarbaugh-Thompson *et al.* (2004) reported that term-limited legislators turn their attention away from constituents and towards interest groups. As we discussed in Section 1, the use of equivalence tests is most helpful to assess 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 re-election incentives are removed is not smaller than the output of their re-election 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 are full-time legislators and electoral competition is high.

3. A research design based on random assignment

A possible strategy to study the effect of removing re-election incentives is to compare the outcomes of legislators who are serving their last term with the outcomes of those legislators who can still run for re-election. This type of observational research design is a common choice (see, for example, Carey *et al.* (1998, 2006), Powell *et al.* (2007) and Sarbaugh-Thompson *et al.* (2004)), and it is often the only design that is 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 behaviour 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—for example, legislators may retire pre-emptively when they expect 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.

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 groups of re-election ineligible and re-election eligible legislators are on average identical at baseline because of 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 4 years and their terms are staggered, with (roughly) half of the 35 senate seats up for election every 2 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 who are elected to a seat in the class of size 18 serve a 2-year term immediately following reapportionment and a 4-year term thereafter, whereas senators who are elected to a seat in the other class serve two successive 4-year terms immediately following redistricting and a 2-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 behaviour. 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 2-year term category and 17 eggs in the 4-year term category, ensuring a fixed margins randomization.

In November 1992, 60% 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 2-year terms and state senators' service to a lifetime maximum of two 4-year terms. An important element of our research design is that 2-year terms do not count against the two-term limit for senators—only 4-year terms do.

Our goal is to measure outcomes for re-election ineligible and re-election 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 re-elected in 1992, and those first elected in 2000 or 2002. Figs 1(a) and 1(b) illustrate the sequence that senators experience based on whether they draw a 2-year or a 4-year term in the 1990s and 2000s respectively. As an example, consider two senators who were elected in November 2002: one assigned a 2-year term and one assigned a 4-year term. Because a 2-year term does not count towards the two-term lifetime limit, the senator who was assigned to serve a 2-year term will stand for re-election in November 2004 and again in November 2008. By contrast, a state senator who was assigned a 4-year lot in 2002 is already on the term limit clock and will stand for re-election only one time in November 2006. In turn, this makes for a legislative session in 2007 (the 86th) where senators who were assigned 4-year lots in 2002 are re-election ineligible whereas senators who were assigned 2-year lots in 2002 still face an election in 2008. The sequence is analogous in the 1990s. However, because of 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.

For analysis, we pool both cohorts, including all senators who were elected in 1992—totalling 35—and all senators who were elected for the first time in 2000 or 2002—totalling 29 (below we explain why we discard six senators). All senators in our treatment group are randomly assigned 4-year terms following the 1990 or 2000 reapportionment and are thus ineligible to run for re-election after the 82nd or 87th legislative sessions. In contrast, all senators in our control group are assigned 2-year terms after the 1990 or 2000 reapportionment and thus are eligible for one more re-election at the end of the 81st or 86th legislative sessions. We study outcomes that are related to legislative output during the 81st and 86th regular sessions of the Arkansas General Assembly (marked in bold in Fig. 1).

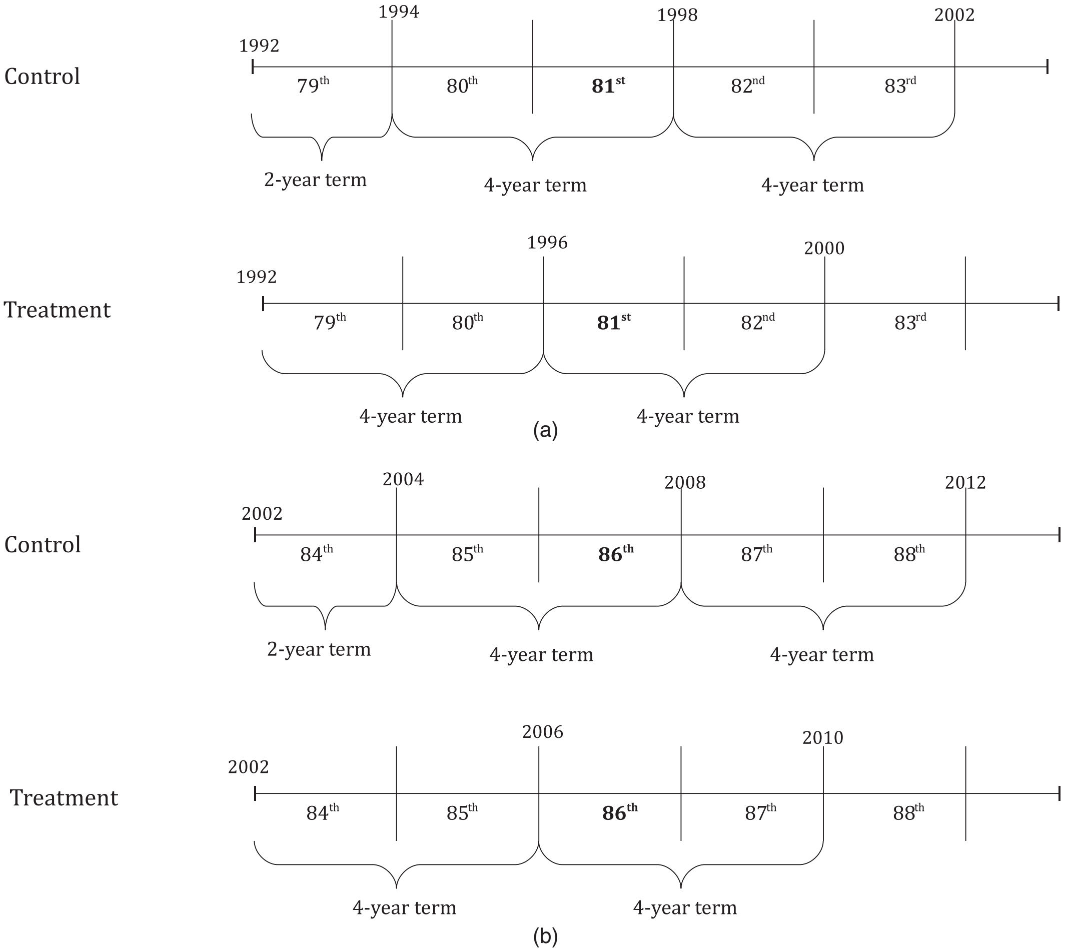

Fig. 1. Illustration of the research design (the numbers in bold (81st and 86th) refer to the legislative sessions in which we examine our outcome): (a) 1990s design; (b) 2000s design

Although some new senators were 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 2-year difference that was induced by the initial random assignment. As illustrated in Fig. 1(a), all Arkansas senators who were elected in November 1992, whether elected for the first time or re-elected, 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 re-elected before 1992, they would all serve their last allowed term at the same time, except for the 2-year discrepancy that was 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 difficult 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 enable 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 we 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 re-election ineligible state and the re-election 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 re-election ineligible and $T_i = 0$ if senator i is re-election 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 re-election 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 re-election ineligible and re-election eligible senators is a valid strategy to learn about the effect of removing re-election incentives.

4.1. 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) and Imbens and Rosenbaum (2005)). For an introduction, see Rosenbaum (2010), section 2, and Imbens and Rubin (2015), section 5; a more advanced treatment can be found in Rosenbaum (2002a). 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 set-up 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, 2002a), 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 $n_1 = 17$ 1s and $n_0 = n - n_1 = 35 - 17 = 18$ 0s. 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 level of significance for a test that rejects H_0 is given by

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

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 the 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 by 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 by 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, 2002a). Using this treatment effect model, we calculate confidence intervals by inverting hypothesis tests, i.e. 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 the 5% level.

Naturally, if the constant treatment effect model is severely misspecified, the confidence intervals that are 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, which was offered by Rosenbaum (2010), section 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 shall 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 (section 2.5 of Rosenbaum (2010), and Rosenbaum (2001)).

Finally, we note that our notation and analysis make the stable unit treatment value assumption (SUTVA). When the SUTVA holds, the outcome of every experimental unit is solely affected by the treatment that is received by that unit, regardless of the treatment status that is assigned to the rest of the units participating in the experiment (see, for example, Rubin (1990) and Bowers *et al.* (2013)). The 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 $\mathbf{t}$ and $\mathbf{t}'$, 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 to 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 the SUTVA because the treatment effect model that we use to impute the missing potential outcomes assumes no interference. Thus, to simplify the exposition, we make the SUTVA throughout.

In our research design, the SUTVA requires that a legislator who is re-election ineligible behave in the same way regardless of how many other legislators in the chamber are re-election ineligible. This would restrict scenarios where, for example, re-election eligible legislators let re-election ineligible legislators have a larger share of those resources that have a fixed budget (e.g. floor time) to help them to take actions that will position them favourably in their quest for higher political office. However, given that re-election 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 co-ordination may be less likely to occur for outcomes that are not directly constrained by the actions of others (e.g. bill introductions).

4.2. Addressing concerns about power: appropriate test statistics and tests of equivalence

To tackle the challenge of low power within the Fisherian framework, we employ both different test statistics 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 0.

4.3. 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, $W_{\mathrm{DM}} = \Sigma_{i:T_i=1} Y_i/n_1 - \Sigma_{i:T_i=0} Y_i/n_0$, which is appealing because of its straightforward average treatment effect interpretation in a Neyman (and also in a superpopulation) framework (Imbens and Rubin, 2015). W_{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 whereas 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), section 5). Robustness to outliers is especially relevant in our study of Arkansas senators, because one of the outcomes that we analyse 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 $W_{\mathrm{DR}} = \Sigma_{i:T_i=1} R_i^y/n_1 - \Sigma_{i:T_i=0} R_i^y/n_0$, 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 (2002a, 2010).

Another strategy to address concerns about statistical power is to employ test statistics based on covariate adjustment (Rosenbaum, 2002b). 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_{DM}, defined as $W_{\mathrm{CDM}} = \Sigma_{i:T_i=1} \hat{e}_i/n_1 - \Sigma_{i:T_i=0} \hat{e}_i/n_0$, where $\hat{e}_i$ is the residual that is obtained from a least squares fit of the observed outcome on K predetermined covariates, $\hat{e}_i = Y_i - \mathbf{x}_i'\hat{\beta}$, with $\mathbf{x}_i$ a $K \times 1$ vector of covariates, and $\hat{\beta}$ 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 $W_{\mathrm{KS}} = \sup_y |\hat{F}_1(y) - \hat{F}_0(y)|$, which measures the maximum absolute difference in the empirical cumulative distribution functions 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 artefact of using differences-in-means and rank-based statistics that may miss certain distributional effects.

Finally, when we analyse covariate balance, we also consider two omnibus test statistics that enable 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 $W_{\mathrm{H}} = \mathbf{d}'\hat{\mathbf{V}}^{-1}\mathbf{d}$, 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 $\hat{\mathbf{V}}$ is the sample variance–covariance matrix of $\mathbf{d}$. The maximum t-statistic is simply $W_{\mathrm{Mt}} = \max(|t_1|, |t_2|, \ldots, |t_K|)$, where $t_k = (\bar{x}_{k1} - \bar{x}_{k0})/\surd(s_{k1}^2/n_1 + s_{k0}^2/n_0)$ 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 that 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.

4.4. Testing the null hypothesis that the treatment effect is non-zero

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 and Hsu, 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 reject the null hypothesis only when there is sufficient evidence that the two groups are similar—i.e. when there is evidence that both $\mu_1 - \mu_0 \geqslant -\delta$ and $\mu_1 - \mu_0 \leqslant \delta$.

We adapt this idea to a Fisherian framework, following related ideas in Rosenbaum (2010), section 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 subhypotheses $H_0^{\delta,+} : y_{1i} - y_{0i} > \delta$ and $H_0^{\delta,-} : y_{1i} - y_{0i} < -\delta$. Given the constant treatment effect model, we have $H_0^\delta : |\tau| \geqslant \delta$, $H_0^{\delta,+} : \tau > \delta$ and $H_0^{\delta,-} : \tau < -\delta$. Our procedure tests both $H_0^{\delta,+}$ and $H_0^{\delta,-}$ by 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 $\{H_0^{\delta,-}, H_0^{\delta,+}\}$ 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 the 5% level—we refer to this value as δ^*. When we find δ^*, we can assert with 95% confidence that the shift τ in legislative output that occurs when a senator goes from being re-election eligible to being re-election ineligible is at most δ^*, i.e. we can assert that $|\tau| \leqslant \delta^*$. Thus, if δ^* is small, we can rule out with 95% confidence that the effect of removing re-election incentives—under a constant treatment effect model—is large. The procedure to find δ^* 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 be rejected. The order is important: we must start with the largest value of δ and subsequently decrease it, to ensure that the sequence $\langle\{H_0^{\delta,-}, H_0^{\delta,+}\}, \delta \in (0, \infty)\rangle$ is a sequentially exclusive partition of hypotheses—i.e. a sequence where, for each value of δ, the set $\{H_0^{\delta,-}, H_0^{\delta,+}\}$ 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 δ^* 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.

5.1. Covariate balance in original experimental sample

Under the random assignment of term length, the ‘treatment effect’ is by construction 0 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 by using the randomization inference approach that was described above. We employ the difference in means between the re-election ineligible and re-election eligible groups as a test statistic W_{DM}, and the two omnibus tests based on W_{H} and W_{Mt} that were described above. We also report tests of equivalence based on the constant treatment effect model. We consider seven predetermined covariates, including the vote share that was obtained in the previous election, party, race, age and gender.

As shown in the second to fifth columns of Table 1, we fail to reject the sharp null hypothesis for every single covariate (the 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_{H} and W_{Mt}, as reported in the lower panel of Table 1. Some of these mean differences, however, are high. For example, 91% of re-election ineligible senators are Democrat and 12% are black, but the corresponding percentages in the re-election 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, four out of 32 senators are black in the treatment group, whereas two out of 32 senators are black in the control group. The difference is only two out of six 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—three of these are in the treatment group and nine in the control group.

Table 1. Covariate balance between re-election ineligible and re-election eligible Arkansas senators, pooling the 81st (1997–1998) and 86th (2007–2008) legislative sessions†

	Means			*Test of no effect p-value*	*Maximum δ failing to reject H_0^{δ}:$\lvert\tau\rvert>\delta$*	
	Treatment group	*Control group*	*Difference*		δ^*	δ^*/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 test				0.3275		
Maximum absolute value t-statistic				0.5394		
Sample size	32	32				

†Treatment group refers to re-election ineligible senators (assigned a 4-year lot in 1992 or 2002), and control group refers to re-election eligible senators (assigned a 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 re-election incentives has no effect for any unit by using the difference-in-means test statistic. Tests of the hypothesis H_0^{δ} that are reported in the last two columns are also randomization based, assuming a constant treatment effect model as explained in the text and employing W_{DM}; sd is the pooled standard deviation across treated and control observations.

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 re-election ineligible and re-election eligible groups are different, using the randomization-based tests of equivalence that were described above. We report the results in the last two columns of Table 1. We report δ^*, the maximum value of δ for which H_0^δ fails to be rejected at 5%, and δ^*/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 δ^*. 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 re-election ineligible senators is shifted no more than ±10.24 percentage points relatively to re-election eligible senators: an absolute difference that represents 0.53 pooled standard deviations, which is a moderately sized effect. Consistent with our findings and discussion above, the values of δ^* 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, on the basis of 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 on-line 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 that is below 5%.

5.2. The effects of removing re-election incentives in the Arkansas Senate

We now study our main question of interest: whether re-election ineligible state senators reduce their legislative output and participation relative to their re-election 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 rate of abstention on roll-call votes and the number of resolutions. For each senator, the rate of abstention 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 that were 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. Although we would prefer a more conventional measure of constituency service (e.g. number of district staff or 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 typical 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 could not 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 2 presents descriptive statistics for all outcome variables.

We note that because of sample attrition our outcome analysis includes fewer observations than the covariate balance analysis that was reported above. By 1997 and 2007, the years when the 81st and 86th legislative sessions began, 15 senators in our sample of 64 had left the chamber: 12 senators left between 1993 and 1997 and three senators left between 2002 and 2007. This leaves us with a remaining sample of 49 senators, whom we call 'survivors'. Any time that 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 length of term that is 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 by using four of the test statistics that were discussed above: W_{DM} (difference in means), W_{CDM} (covariate-adjusted difference in means), W_{DR} (difference in average ranks) and W_{KS} (the maximum distance between the empirical cumulative distribution functions). In addition, we present three different confidence intervals employing the test statistics W_{DM}, W_{CDM} and W_{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 on-line supplemental appendix, we present additional results based on the covariate-adjusted KS statistic and the difference in interquartile ranges for the pooled sample, and also separate 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 that 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. On the basis of this pattern, there is not enough empirical evidence to assert that the removal of re-election incentives affected legislative participation in the Arkansas Senate during the sessions that we study.

But there are further lessons to be learned from our results. The confidence interval for the rate of abstention based on the difference in means W_{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 that were reported in Table 2, where the rate of abstention in the re-election ineligible group (2.36%) is shown to be more than twice the rate in the re-election eligible group. However, as illustrated in Fig. 2, this mean difference is driven entirely by one senator in the treatment group, Steve Faris, who missed approximately 31% 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 that are based on this statistic (and on its covariate-adjusted version W_{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_{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_{DR} is unaffected by the extreme value of the rate of abstention that is exhibited by Faris.

Moreover, covariate adjustment is shown to reduce the length of confidence intervals modestly in some cases. To compute the covariate-adjusted difference in means statistic W_{CDM}, we employ

Table 2. Descriptive statistics for outcome variables in the Arkansas Senate, pooling 81st (1997–1998) and 86th (2007–2008) legislative sessions†

	Minimum		*Maximum*		*Mean*		*Median*		*sd*		*1st quartile*		*3rd quartile*	
	Treatment group	*Control group*	*Treatment group*	*Control group*	*Treatment group*	*Control group*	*Treatment group*	*Control group*	*Treatment group*	*Control group*	*Treatment group*	*Control group*	*Treatment group*	*Control group*
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

†Treatment group refers to re-election ineligible senators (assigned a 4-year lot in 1992 or 2002), and control group refers to re-election eligible senators (assigned a 2-year lot in 1992 or 2002). The number of observations is 26 in the treatment group and 23 in the control group, except for bills cosponsored, which include 12 treated and 13 control observations.

Table 3. Test of a sharp null hypothesis and confidence intervals based on different test statistics for outcome variables in the Arkansas Senate, pooling the 81st (1997–1998) and 86th (2007–2008) legislative sessions†

	p-value from test of sharp null				*95% confidence interval for constant effect*		
	W_{DM}	W_{CDM}	W_{DR}	W_{KS}	W_{DM}	W_{CDM}	W_{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	$[-14.26, 6.16]$	$[-14.16, 6.11]$	$[-15.99, 5.99]$

†p-values correspond to a 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 re-election incentives, and the tests are based on a comparison of re-election ineligible senators (assigned a 4-year lot in 1992 or 2002) and re-election eligible senators (assigned a 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.

the residuals from a least squares fit of each outcome on previous vote share. For bills introduced, employing W_{CDM} instead of W_{DM} leads to a reduction of 4.5% 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_{CDM} and W_{DM} is very similar.

Finally, the fact that we reach the same conclusions when we employ W_{DM}, W_{CDM} and W_{DR} as when we employ the KS statistic W_{KS} suggests that our failure to reject the sharp null hypothesis by 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 re-election ineligible and re-election eligible groups are statistically indistinguishable.

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

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 because of 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 re-election is equal to his or her output under re-election 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.

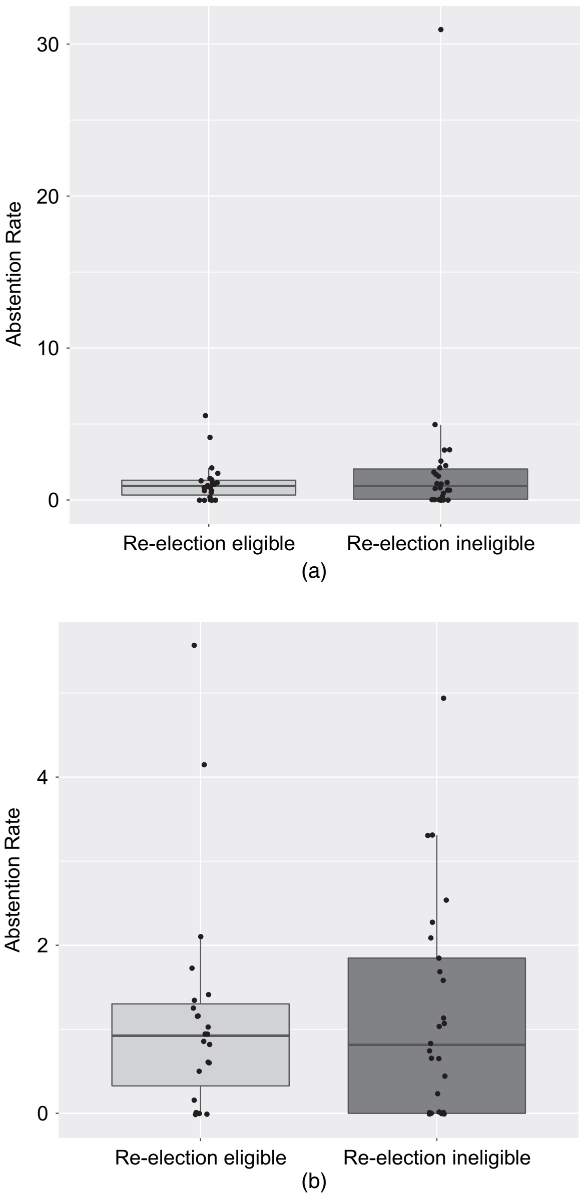

Fig. 2. Removal of re-election effects on abstention rates in the Arkansas Senate—81st (1997–1998) and 86th (2007–2008) legislative sessions: (a) abstention rate; (b) abstention rate without Steve Faris

The second to fourth columns of Table 4 report δ^*, the maximum value of τ that fails to reject the null hypothesis of non-equivalence, H_0^δ. We present results based on two statistics: W_{DM} and W_{DR}. As in Table 1, we report both δ^* and δ^*/sd for each test. On the basis of the results for W_{DM}, we can say with 95% confidence that $|\tau|$ is at most approximately six bill introductions, four bills passed, an abstention rate of 3%, 1.5 resolutions and 13 bills cosponsored.

The results employing the rank-based statistic W_{DR} lead to generally similar conclusions, with some important exceptions. For abstentions, δ^* decreases dramatically from 3.08 to 0.65,

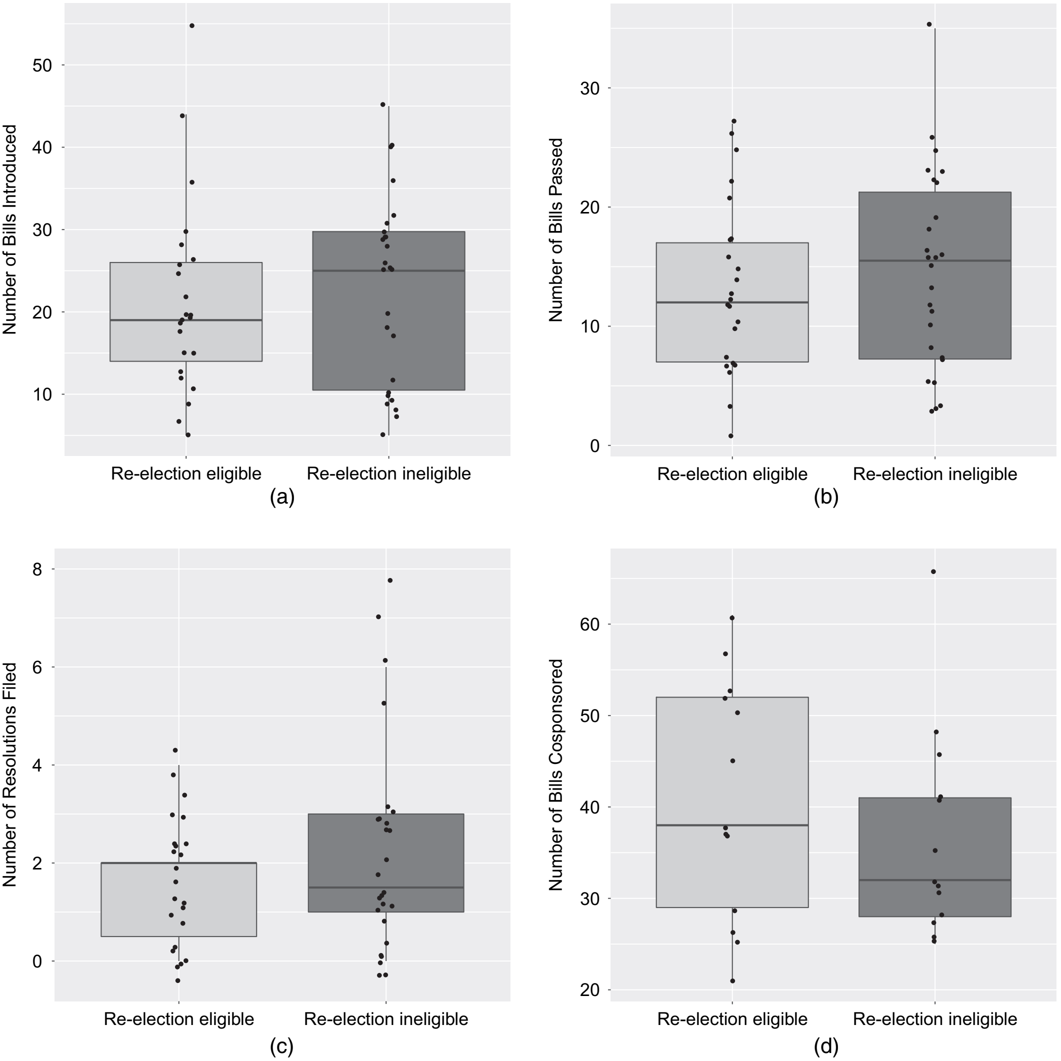

Fig. 3. Removal of re-election effects for bill introduction, passage and symbolic bills in the Arkansas Senate—81st (1997–1998) and 86th (2007–2008) legislative sessions: (a) bills introduced; (b) bills passed; (c) resolutions filed; (d) bills cosponsored

illustrating once again that the conclusions that are 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 that are based on W_{DM} and W_{DR} are very similar when the outlier is excluded.

For resolutions, the rank-based value of δ^* also decreases considerably relative to the value that is based on W_{DM}, from 1.5 to 1, but the results for bills introduced lead to a larger δ^*. In general, δ^* ranges from a small effect of 0.15 standard deviations in the case of the rate of abstention, 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 δ^*/sd for abstentions by using all observations is higher if we use the standard deviation of the control group instead. This is because the standard

Table 4. Tests of equivalence and negative effects (shirking), pooling the 81st (1997–1998) and 86th (2007–2008) legislative sessions†

	Maximum δ failing to reject H_0^{δ}:$\vert\tau\vert > \delta$				*Maximum δ failing to reject $H_0^{\delta,\mathrm{S}}$:$\tau < -\delta$*			
	W_{DM}		W_{DR}		W_{DM}		W_{DR}	
	δ^*	δ^*/sd	δ^*	δ^*/sd	δ^*	δ^*/sd	δ^*	δ^*/sd
Abstentions	3.08	0.69	0.65	0.15	3.1	0.7	0.65	0.15
Abstentions without Steve Faris	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						

†Tests of the hypotheses H_0^{δ} and $H_0^{\delta,\mathrm{S}}$ are performed by 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 re-election incentives, and the tests are based on a comparison of re-election ineligible senators (assigned a 4-year lot in 1992 or 2002) and re-election eligible senators (assigned a 2-year lot in 1992 or 2002). Calculations for rates of abstention excluding Steve Faris (in the re-election 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 $H_0^{\delta,\mathrm{S}}:\tau \geqslant \delta$. The number of observations for bills cosponsored is 15 treated and 14 control.

deviation in the treatment group is considerably higher because of the outlier observation. For example, for the results that are based on W_{DR}, $\delta^*/\mathrm{sd} = 0.15$ 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 that are associated with removing re-election incentives. As mentioned in Section 1, 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 re-election ineligible senators participating at equivalent or higher rates, not lower, for many of the outcomes that 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 re-election ineligible than if he is re-election eligible—i.e. we test the null hypothesis of shirking, $H_0^{\delta,\mathrm{S}}:\tau < -\delta$, for $\delta > 0$. Note that $\tau = y_{i1} - y_{i0} < 0$ implies shirking only for the bill and resolution outcomes, but not for our rates of abstention, since a *positive* τ for abstentions is a shirking effect. For this reason, for this outcome we define the null hypothesis of shirking as $H_0^{\delta,\mathrm{S}}:\tau > \delta$, for $\delta > 0$.

The results, which we report in the sixth to ninth columns 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 enable us to rule out large negative effects of removing re-election incentives on legislative output and participation (under a constant treatment effect model). The results show that, in several cases, the δ^* for the 'shirking' hypothesis $H_0^{\delta,\mathrm{S}}$ is considerably smaller than the δ^* for the two-sided hypothesis of non-equivalence $H_0^{\delta}:|\tau| > \delta$. For example, the results for bills introduced based on W_{DR} show that δ^* decreases from 6.99

to 3.98 when we switch from H_0^δ to $H_0^{\delta,\mathrm{S}}$, enabling us to assert that $\tau \geqslant -3.98$ and thus to rule out negative effects that could not be ruled out on the basis of a test of H_0^δ, which could only justify the assertion that $\tau \geqslant -6.99$. This is an improvement, but we note that, although four bill introductions are 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 δ^* that occurs when considering $H_0^{\delta,\mathrm{S}}$ instead of H_0^δ is very large: from 1.5 to 0.25 when using W_{DM}, and from 1 to 0 when using W_{DR}. The results for resolutions reflect a phenomenon that is illustrated in Table 2 and Fig. 3(c): senators in the re-election ineligible group tend to introduce more resolutions than senators in the re-election 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 re-election 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, which is reported in the last row. Because the re-election ineligible group has a lower average of cosponsored bills than the re-election eligible group (see the last row in Table 2), the δ^* that is associated with the two-sided H_0^δ is roughly the same as the δ^* that is associated with the negative effect one-sided $H_0^{\delta,\mathrm{S}}$. 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 δ^*. But this will not occur when treated responses tend to be lower than control responses, as occurs for bills cosponsored.

5.3. 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 rates of attrition 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 re-election ineligible group and 23 senators in the re-election eligible group. The difference in the non-missingness rates by group (23/32 and 26/32) are statistically indistinguishable. This balance in rates of attrition is also seen when we consider each legislative session individually. In the 1990s cohort, eight senators who were assigned 2-year terms and four senators who were assigned 4-year terms drop out of the sample before 1997 (the null hypothesis that the true probability of success is 0.5 in 12 Bernoulli trials is not rejected, p-value 0.3877), and in the 2000s cohort one senator who was assigned a 2-year term and two senators who were assigned 4-year terms drop out of the sample before 2007.

We also note that, in addition to the initial randomization, which enables 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 re-election is likely to affect both groups equally, and the composition of both groups in terms of ‘departers’ (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 because of the initial randomization.

Table 5. Covariate balance between re-election ineligible and re-election eligible Arkansas Senators, pooling the 81st (1997–1998) and 86th (2007–2008) legislative sessions—only survivors†

	Means			*Test of no effect p-value*	*Maximum δ failing to reject H_0^δ:$\|\tau\| > \delta$*	
	Treatment group	*Control group*	*Difference*		δ^*	δ^*/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 test				0.1003		
Maximum absolute value t-statistic				0.3746		
Sample size	26	23				

†Treatment group refers to surviving re-election ineligible senators (assigned a 4-year lot in 1992 or 2002), and control group refers to surviving re-election eligible senators (assigned a 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 re-election incentives has no effect for any unit using the difference-in-means test statistic. Tests of the hypothesis H_0^δ that are reported in the last two columns are also randomization based, assuming a constant treatment effect model as explained in the text and employing W_{DM}; sd is the pooled standard deviation across treated and control observations.

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 re-election ineligible group who survived until their last term may be different from senators in the re-election 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 δ^* 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 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 departers—where the share of observations in each subgroup is known but the outcomes in the departer group are not.

To calculate the upper bound on the average treatment effect, we set the outcome values of those senators who were 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 who were initially assigned to the

Table 6. 75th- and 25th-percentile bounds on the average treatment effect of re-election ineligibility in the Arkansas Senate, pooling the 81st (1997–1998) and 86th (2007–2008) legislative sessions†

	Average treatment effect lower bound	*Average treatment effect upper bound*
Abstentions	0.65	1.37
Abstentions without Steve Faris	−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

†The columns labelled 'Average treatment effect lower bound' and 'Average treatment effect upper bound' report respectively the estimated lower and upper bounds of the average treatment effect—the average potential outcome under the re-election ineligibility (treated) condition minus the average potential outcome under the re-election eligibility (control) condition. Upper bounds set missing treated outcomes to the 75th percentile of observed treated outcome and missing control outcomes to the 25th percentile of observed control outcomes. The lower bound is analogous, using the 25th percentile for missing treated and the 75th percentile for missing control. Calculations for abstention rates excluding Steve Faris (in the treatment group) use a total sample size of 63; the number of observations for bills cosponsored is 15 treated and 14 control.

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 re-election ineligible group, if observed, would have been low whereas all the missing outcomes in the re-election 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 whereas 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 re-election minus the average potential outcome under the possibility of one more re-election—for each outcome. Since our intention is to assess the robustness of our conclusion that removing re-election 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 rates of abstention. The results in Table 6 show that, for bill introductions, bill passage, resolutions and rates of abstention, 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 re-election ineligible senators would have introduced just 12 bills (25th percentile) whereas all missing re-election 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 rates of abstention. The lower bound on the last-term effects on bill passage is just −1.78 bills, and this is assuming that missing re-election ineligible senators would have passed just seven bills whereas missing re-election 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 rates of abstention among re-election ineligible senators would have been 1.68% whereas the rate of abstention among missing re-election eligible senators would have been 0.16%: about a tenth the size. (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 re-election 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 or sampling uncertainty.

6. Conclusion

We examine how removing re-election incentives affects legislative behaviour in the Arkansas Senate, specifically the extent to which re-election ineligible legislators produce less legislation and abstain at higher rates than their re-election 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 re-election 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 re-election 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 departure 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 ‘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. re-election ineligibility leads to less legislative output and higher abstentions. This hypothesis of last-term shirking has been at the centre of prior studies.

We find that one-sided tests are more informative and for most outcomes enable 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 enable 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 that 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 re-election eligible and re-election 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 that 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 on the basis of these results alone that the negative effects of removing re-election 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 re-election incentives when compared with the possibility of being re-elected 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 that 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 non-parametric way. We believe that 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.

Acknowledgements

We thank Steve Cook, 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 Joint Editor, the Associate Editor, two 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.

References

Berger, R. L. and Hsu, J. C. (1996) Bioequivalence trials, intersection-union tests and equivalence confidence sets. *Statist. Sci.*, **11**, 283–319.

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

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

Carey, J., Niemi, R., Powell, L. and Moncrief, G. (2006) The effects of term limits on state legislatures: a new survey of the 50 states. *Legisltv. Stud. Q.*, **31**, 105–134.

Cattaneo, M. D., Frandsen, B. and Titiunik, R. (2015) Randomization inference in the regression discontinuity design: an application to party advantages in the U.S. senate. *J. Causl Inf.*, **3**, 1–24.

CQ Press (2016) *Voting and Elections Collection*. New York: Sage.

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

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

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

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

Gerber, B. J. and Teske, P. (2000) Regulatory policymaking in the American states: a review of theories and evidence. *Polit. Res. Q.*, **53**, 849–886.

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

Ho, D. E. and Imai, K. (2006) Randomization inference with natural experiments. *J. Am. Statist. Ass.*, **101**, 888–900.

Imbens, G. W. and Rosenbaum, P. R. (2005) Robust, accurate confidence intervals with a weak instrument: quarter of birth and education. *J. R. Statist. Soc.* A, **168**, 109–126.

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

Mansbridge, J. (2009) A selection model of representation. *J. Polit. Philos.*, **17**, 369–398.

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

Manski, C. F. (2007) *Identification for Prediction and Decision*. Cambridge: Harvard University Press.

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

National Conference of State Legislatures (2009) Full and part-time legislatures. National Conference on State Legislature, Washington DC.

Nicholson-Crotty, S. (2009) The politics of diffusion: public policy in the American states. *J. Polit.*, **71**, 192–205.

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

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

Rosenbaum, P. R. (2002a) *Observational Studies*, 2nd edn. New York: Springer.

Rosenbaum, P. R. (2002b) Covariance adjustment in randomized experiments and observational studies. *Statist. Sci.*, **17**, 286–327.

Rosenbaum, P. R. (2007) Interference between units in randomized experiments. *J. Am. Statist. Ass.*, **102**, 191–200.

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

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

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

Sarbaugh-Thompson, M. (2010) Measuring 'term limitedness' in US multi-state research. *State Polit. Poly Q.*, **10**, 199–217.

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

Sekhon, J. and Titiunik, R. (2012) When natural experiments are neither natural nor experiments. *Am. Polit. Sci. Rev.*, **106**, 35–57.

Shipan, C. R. and Volden, C. (2006) Bottom-up federalism: the diffusion of antismoking policies from US cities to states. *Am. J. Polit. Sci.*, **50**, 825–843.

Squire, P. (2007) Measuring state legislative professionalism: the Squire index revisited. *State Polit. Poly Q.*, **7**, 211–227.

Titiunik, R. (2016) Drawing your senator from a jar: term length and legislative behaviour. *Polit. Sci. Res. Meth.*, **4**, 293–316.

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

Supporting information

Additional ‘supporting information’ may be found in the on-line version of this article:

‘Supplemental online appendix to “Legislative behavior absent reelection incentives: findings from a natural experiment in the Arkansas Senate”’.