Leadership or Luck?
Randomization Inference for Leader Effects (RIFLE)

Christopher R. Berry and Anthony Fowler[1]
crberry@uchicago.edu; anthony.fowler@uchicago.edu
Harris School of Public Policy
University of Chicago

First Draft: February 10, 2017
This Draft: November 15, 2017

Abstract

Anecdotal evidence suggests that some political leaders are more effective than others, causing better outcomes for their citizens. However, observed differences in outcomes between leaders could be attributable to chance variation. We develop RIFLE, a quantitative test of leader effects, and implement it for world leaders, U.S. governors, and U.S. mayors and for several outcomes. We find clear evidence that world leaders matter for GDP growth and some evidence that U.S. governors matter for crime and public finance, but we also obtain several surprising null results. Our test can be applied to virtually any setting with leaders and an objective outcome of interest, so its continued application should improve our understanding of where, when, and why leaders matter.

[1] Authors contributed equally. We thank Mia Greco, Satyen Gupta, and Johnathan Guy for research assistance, and we thank Scott Ashworth, Ethan Bueno de Mesquita, Steven Durlauf, Andy Eggers, Alex Fouirnaies, Justin Grimmer, Evan Haglund, Will Howell, Guillaume Pouliot, Jas Sekhon, and seminar participants at Chicago, LSE, MPSA, NYU, and Warwick for helpful comments.

The U.S. economy grew at an annual rate of over 6 percent during Harry Truman's second term, faster than under any other postwar president. George H. W. Bush presided over an economy growing at less than 1 percent on average during his second term, the worst record over the same period (Blinder and Watson 2016). Under the administration of Mayor Rudolph Giuliani, violent crime fell by more than 56 percent in New York City in 1990's, compared to a decline of 28 percent nationwide (Corman and Mocan 2005). By contrast, under the leadership of Rahm Emmanuel, homicides have increased 70 percent since 2014 in Chicago, which was the only major U.S. city to experience an increase in murders in both 2015 and 2016 (Kennedy and Abt 2016; Sanburn 2016).

How much credit do leaders such as Truman and Giuliani deserve for the outcomes that happened on their watch? How much blame should fall at the feet of those like Bush and Emmanuel? Do the decisions and actions of leaders change the course of events, or are some merely (un)lucky, holding office at a time when other factors would have generated largely the same outcomes regardless of who sat behind their desk? This question—whether leaders matter—has fascinated scholars for centuries. Today, colleges and universities offer advanced degrees in leadership, and airport bookstores feature bestsellers on the topic, implying a settled conclusion that leaders matter. Yet there is little rigorous empirical evidence on the effects of leaders on outcomes of interest to social scientists.

In this paper, we aim to make two contributions to the study of leadership. First, we introduce a new method for statistically testing leader effects, which has several advantages relative to other methods that have been used in the literature. We call this method Randomization Inference for Leader Effects or RIFLE. Second, whereas the extant literature has

focused on assessing the effects of national leaders on the economy, we extend our analysis to include subnational leaders and non-economic outcomes.

Related Literature

The subject of leadership has fascinated scholars at least back to the ancient Greeks. And for just as long, scholars have debated whether leaders matter in shaping outcomes for better or worse. Thucydides chronicled the great leaders, such as Pericles, whose particular decisions and abilities, in his view, determined the outcome of the Peloponnesian War. Meanwhile, Plato extolled the virtues of the philosopher-king while advocating a system of education and selection that would make any individual leader replaceable by another who would make similar decisions.

In the nineteenth century, Thomas Carlyle (1859) advanced the still-influential view that "The history of the world is but the biography of great men." But his contemporary, Karl Marx (1852), argued that historically determined social and economic forces constrain the choices available to leaders, making any individual relatively unimportant in the course of events.

More recently, the role of leadership has been overshadowed in political science and economics by the role of institutions in determining the fate of nations. Building on the work of Douglas North (1990), a dominant theme in the contemporary literature on economic development is that good institutions are the fundamental cause of long-run economic growth (e.g., Acemoglu, Johnson, and Robinson 2005), although empirically identifying the effects of institutions is not unproblematic (e.g., Glaeser et al. 2004).

While institutions have taken center stage in contemporary scholarship on economic development, leaders have not been entirely forgotten. An emerging formal literature tackles

fundamental questions about the nature and source of leadership (see Levi and Ahlquist 2011). Notable contributions include Dewan and Squintani (2015) on relations between a leader and followers, Canes-Wrone (2006) on the distinction between leadership and pandering, and Dewan and Myatt (2008) on leadership and communication.

From an empirical perspective, a seminal paper by Jones and Olken (2005) rekindled interest in the study of leadership by providing the most credible evidence to date that political leaders matter. They use unexpected deaths of world leaders while in office as a source of exogenous variation in leadership. They show that the rate of economic growth in a country changes when a leader dies in office. Their key empirical test is whether the change in economic growth between the last two years of one leader's tenure and the first two years under the successor is greater than would be expected by chance. They find strong evidence of abnormal variation in growth around exogenous leader transitions, which implies that leaders matter. They show further evidence that unexpected leader turnover leads to changes in economic growth in autocracies but not democracies.

The Jones and Olken findings have been extended in several ways by subsequent authors. Besley et al. (2011) find that educated leaders particularly exert a positive effect on economic growth. Within the Jones-Olken framework, they show that a transition from a more educated to a less educated leader results in a reduction in economic growth, while a transition from a less educated to a more educated leader leads to a boost in economic growth. Yao and Zhang (2015) analyze the effects of city leaders in China on local economic growth, taking advantage of the fact that leaders regularly move, so that the same person may be mayor of multiple cities over the course of her career. Yao and Zhang use all leader transitions in their analysis rather than identifying unexpected transitions as in Jones and Olken, and they find mixed results depending

on the statistical test they consider. Easterly and Pennings (2016) challenge Jones and Olken's conclusion that leaders matter more in autocracies than in democracies. Specifically, Easterly and Pennings estimate leader fixed effects in a growth model and show that the variance of the fixed effects is at least as large in democracies as in autocracies. While there is more total variability in growth rates in autocracies, they contend that the amount of variation attributable to leaders is higher in democracies.

Randomization Inference for Leader Effects (RIFLE)

Our goal is to test whether leaders matter for particular outcomes of interest. As with the previous literature, we cannot estimate the effects of any individual leader with statistical confidence, but we can ask whether leaders matter in the aggregate. Are some leaders better than others, such that we can statistically reject the null hypothesis that all leaders are the same with respect for a particular outcome? After introducing our method, we compare it with other methods used in the prior literature and explain how it can improve or complement them.

There are several methodological challenges associated with estimating leader effects. A key challenge, and the primary focus of this paper, is inference. Suppose we observe some apparent correlation between leaders and outcomes as in the examples discussed in the introduction. We would expect some of that apparent correlation just by chance, and we'd like to account for the idiosyncrasies of luck to determine whether that correlation is indeed statistically significant. As we'll show, because of random noise and serial correlation in leaders and outcomes, existing methods can fail to distinguish leadership from luck.

Another set of challenges has to do with identification. If we detect a statistically significant relationship between leaders and outcomes, it might be attributable to leader effects or

there might be other reasons that outcomes systematically correspond with leaders. For example, if the outcome of interest affects leader transitions, this could also generate a correlation between leaders and outcomes. Although we cannot entirely remove these concerns, we can show through simulations and sensitivity analyses that the substantive relevance of these identification concerns is minimal.

Our general strategy involves regressing an outcome on leader fixed effects, recording a summary statistic of fit, and then simulating the distribution of summary statistics that we would expect under the null. As summary statistics of fit, the r-squared, adjusted r-squared, and F-statistic will all produce identical p-values and implied effect sizes in our subsequent analyses because for a given sample size and number of regressors, these statistics all increase monotonically as the others increase. For the purposes of this paper, we focus on the r-squared statistic, which is familiar to social scientists and has a substantive interpretation as the proportion of variation in the outcome that is explained by the leader fixed effects. However, if subsequent practitioners would prefer using another fit statistic, that is perfectly allowable within our framework.

In and of itself, the r-squared statistic is not particularly informative. A high value could reflect leader effects, but it could also reflect within-unit variation over time unrelated to leader effects, or it could suggest that the regression with many independent variables over fit random variation in the outcome. Therefore, we need a strategy for simulating the distribution of r-squared statistics that we would expect under the null hypothesis of no leader effects. To do this, we randomly permute the ordering of leaders within each unit, keeping the tenure of each leader the same as in the real data set but varying the order in which each leader served. For each random permutation, we again regress growth on leader fixed effects and record the r-squared

statistic. We repeat this procedure many times to estimate the distribution of r-squared statistics we would obtain under the null of no leader effects. The proportion of random permutations that produce an r-squared statistic greater than that from the real data is an estimated p-value testing the sharp null hypothesis of no leader effects.[2]

Prior to implementing our method, we take several steps to prepare the data for analysis. These steps are optional but can improve statistical precision. We start by converting each outcome variable to proportionate growth by subtracting the lagged value and then dividing by the lagged value. This accounts for variation in the level of each outcome across geographic units. Next, we de-mean growth by year to remove consistent time trends across units.[3] Then, if there are any observations with missing data, we drop those observations and stitch together blocks of time, coding a new time variable so that we have a single, contiguous stretch of time with complete data for each unit. This final step allows us to take advantage of all the observations with non-missing values when permuting the data.

Summary of our Procedure
1. De-mean the outcome by year to remove consistent time trends across units (optional).
2. Drop missing observations and "stitch" together valid observations for each unit (optional).
3. Regress the outcome on leader fixed effects and record the r-squared statistic.
4. Randomly permute leaders within each unit, sampling each leader stint as a block.
5. Again, regress growth on leader fixed effects and record the r-squared statistic.
6. Repeat steps 4 and 5 many times, recording the proportion of cases where the r-squared from the permuted data is greater than that from the real data.

Our procedure is closely related to many others utilizing some form of permutation test or randomization inference (e.g., Abadie, Diamond, and Hainmueller 2010; Fisher 1935; Ho and

[2] These hypothesis tests are one sided because there is no reason to expect the real r-squared statistic to be smaller than the expected r-squared statistic under the null. If some leaders are better than others, this will only increase the value of the real r-squared statistic.

[3] We could also de-mean by unit but this would have no impact on our subsequent p-values. Our inferential strategy implicitly accounts for variation in growth across units.

Imai 2006; Rosenbaum 2002). One distinction between our approach and a canonical application is that we are not estimating a treatment effect directly. Instead, we utilize permutations tests in order to estimate the distribution of r-squared statistics that we would expect under a null hypothesis of no leader effects. Our approach is also closely related to methods for testing hypotheses about multiple treatments (e.g., Wooldridge 2010, pp. 963-967). One could think of each leader as a separate treatment, where we want to test whether the treatment effects are jointly distinguishable from one another. However, rather than using conventional approaches, we must utilize blocked randomization inference in order to account for time trends and serial correlation.

The logic of our random permutation tests is as follows. Assume that leader transitions are unrelated to potential outcomes, such that in the absence of any leader effects, there should be no systematic correspondence between leaders and outcomes. There are three ways we can get a high r-squared statistic when we regress growth on leader indicators. First, there could be leader effects, and this is what we'd like to identify. Second, there could be serial correlation or genuine trends in growth over time within units even in the absence of leader effects, and the leaders who happened to serve in good times will get credit for this in the regression. Third, the leader fixed effects could be over-fit to random, year-to-year fluctuations in growth or even measurement error, further inflating the r-squared statistic. Therefore, in order to test for leader effects, we'd like our random permutation tests to incorporate the last two factors but not (all of) the first.

In our random permutations, the number of fixed effects in each regression is held constant, and the distribution of tenure across leaders is also held constant. This means that the

extent of overfitting is the same, in expectation, in the real data and the permuted data.[4] Furthermore, if there is serial correlation or unit-specific time trends unrelated to leaders, this will inflate the r-squared from the permuted regressions in the same way, in expectation, as it inflates the r-squared from the real regression. In either case, some leaders might wrongly receive credit for good times. However, if there are genuine leader effects, this will increase the r-squared in the real regression by more than it increases the r-squared in the permuted regressions. Therefore, if the r-squared from the real data is larger than that from the random permutations, this is an indication that ebbs and flows in growth coincide with the intervals of time in which different leaders served, suggesting that some portion of that r-squared statistic can be attributed to leader effects rather than just serial correlation or chance.

Our practice of sampling each leader as a block and maintaining the same distribution of contiguous periods of service in our permutations is important. To our knowledge, the approach in the literature closest to our own is a robustness check from Yao and Zhang (2015, p. 420). However, instead of sampling each leader as a block, Yao and Zhang sample each year independently such that leaders' terms are no longer contiguous in the random permutations. This approach accounts for the possibility of overfitting discussed above, but it does not account for the possibility of serial correlation or unit-specific time trends, and as a result, this test is likely to reject the null even if there are no leader effects.

To understand how our method performs in different scenarios, let's consider how varying features of the data generating process will influence the r-squared statistic in the real

[4] This assumes that the researcher does not use the observed data to make specification choices. If a careless researcher modified the above procedure to better fit the observed data, the resulting p-values would be misleading. This is, of course, a concern with virtually all quantitative analyses, although we attempt to mitigate these concerns in this case by specifying a simple and generalizable procedure that will be applied in the same way to different data sets.

and permuted data sets. Recall that all of the regressions run under RIFLE will include a set of leader fixed effects. By definition, $r^2 \equiv 1 - \frac{RSS}{TSS}$, where RSS is the residual sum of squares and TSS is the total sum of squares. In our regressions, the TSS is identical for both the real data and the permuted data sets where the ordering of leaders is randomly shuffled. Therefore, to think about how our method works, we need to think about how leader effects, time effects, and random noise influence the RSS. Random noise increases the RSS, and it increases the RSS in the same way, in expectation, in the real data set and the permuted data sets. If the noise was expected to affect the RSS differently in the real data, it wouldn't be random. Similarly, time effects that are unrelated to leader tenures will also increase the RSS, and they will increase the RSS the same way, in expectation, in the real and the permuted data sets. This is why our method of permuting leader tenures accounts for noise and time trends unrelated to leaders.

How do leader effects influence the RSS? A constant effect for each leader should have no effect on the RSS in the real data set. In this context, $RSS \equiv \sum_i \sum_t (Y_{it} - \bar{Y}_i)^2$, where i denotes leaders and t denotes observations within each leader. In other words, the RSS is the sum of squared deviations of each data point from the mean for each leader. A constant effect for each leader would mean that the outcome is shifted by the same amount for all observations within each leader, such that each Y_{it} would be shifted by the same amount as each $\bar{Y}_i$, and the RSS would be unchanged by leader effects. However, in the permuted data sets, leader effects would increase the RSS. For each permuted leader tenure that overlaps with multiple actual leader tenures, leader effects will shift observations by different amounts within each permuted leader, thereby increasing the RSS. This means that in the presence of genuine leader effects, we expect the RSS to be lower for the real data set than in the permuted data sets, meaning that the r-squared will be higher.

To fix ideas, consider the simplest possible example where our test would allow us to say something about leader effects. Suppose there is 1 unit with 2 leaders and 3 periods. Without loss of generality, suppose Leader A served during the first two periods, and Leader B served during the last period. In this simple example, there are only two ways to permute the leaders. We can assign Leader A to the first two periods—as in the real world, or we can assign her to the last two periods. If Leader A is better than Leader B, or vice versa, we would expect the outcome from the first two periods to be more similar to each other than they are to the value from the third period, and the real data will give a higher r-squared statistic. If there are no leader effects but there is random noise or serial correlation, either permutation is equally likely to give a higher r-squared.

This simple example illustrates several features and limitations our approach. First, identification comes from leaders who serve different periods of time. If there were 4 periods and each leader served two periods, both permutations would yield the same r-squared. Next, our procedure behaves poorly when there are few leaders per unit. In the example above, the p-value can only take one of two possible values, but asymptotic refinement improves quickly with more leaders, so long as there is variation in lengths of service. Furthermore, our procedure requires that we put some structure on the timing of leader effects. Suppose a leader's policy decisions primarily affect growth in the next year. In that case, we would have no opportunity to detect leader effects in the simple example above. For our subsequent analyses, we will assume that leaders can only affect outcomes in the years in which they serve, but one could easily conduct separate tests where they assume that these effects are lagged by one year, two years, etc. Lastly, our approach does not require us to hypothesize that one particular leader is better than another.

For the purposes of this study, we are agnostic about which leaders are better. We test whether some leaders are different from others in ways that matter for various outcomes of interest.

RIFLE is a general, flexible method that can be applied to any setting with an objective outcome of interest and leaders who serve different periods of time. We have developed a Stata package that will allow future researchers to easily apply this method to many different contexts and outcomes in order to better understand where, when, and why leaders matter.

Comparison with Other Methods

RIFLE resembles methods used in other papers that rely on leader fixed effects in one way or another (e.g., Bertrand and Schoar 2003; Easterly and Pennings 2016). Our approach to inference is different, however. We do not assume that the r-squared from a set of leader-specific fixed effects will be zero when leaders do not actually affect the outcome. Because of random noise in the data, serial correlation, and unit-specific trends, leader fixed effects will contribute to the r-squared even when leaders do not matter. Using the adjusted r-squared, as Bertrand and Schoar (2003) do, partially accounts for the problem of overfitting to random noise, but it does not address time trends and serial correlation. RIFLE accounts for all these factors, without requiring additional assumptions about the nature of the serial correlation or the unit-level trends. In this sense, our approach is more conservative and will be less prone to detecting leader effects in cases where there are none.

A preview of one of our results illustrates the idea that the r-squared or adjusted r-squared statistics alone are insufficient for assessing leader effects. In a subsequent analysis, we test whether U.S. mayors affect employment in their cities. When we randomly permute the terms of service for mayors, there should be no leader effects because terms of service are random.

Nonetheless, the average r-squared statistic from our random permutations in this setting is .606, meaning 61 percent of the variation in employment across cities appears to be explained by leader fixed effects, even though the leader variables correspond to random times. The adjusted r-squared, which is designed to mitigate overfitting is still .525. If we first de-mean by city, the r-squared and adjusted r-squared statistics are .343 and .209, respectively. Clearly, none of these statistics can separate leader effects from other forces because the numbers are large even when there are no leader effects. As we'll see, although the r-squared and adjusted r-squared statistics are large when we regress city employment on mayor fixed effects, they are no larger than we would expect by chance if mayors do not affect employment.

RIFLE also offers some advantages relative to the method of Jones and Olken (2005). They use a similarly nonparametric inferential approach, comparing the changes in economic growth in periods with leadership transitions to the distribution of changes in periods when there are no transitions. However, their analysis only includes leader transitions arising from unexpected deaths—i.e., those due to accident of illness in office. While unexpected deaths provide plausibly exogenous changes in leadership, identification comes from a relatively small number of leader transitions. Specifically, they have only 57 such unexpected transitions from a panel of 130 countries since 1945. Our approach utilizes information from virtually all leaders and all countries.

A concern with focusing on unexpected leader deaths is that the resulting changes in economic growth may also reflect the disruptive effects of an unanticipated transition of power. To the extent that unexpectedly changing leaders disrupts the government or economy, this will be reflected in the Jones and Olken estimates. While they take measures to reduce the possibility that disruption contaminates their results—such as excluding the first year or two after the

transition—some concerns remain. Besley et al. (2011) show that the average change in growth following an unexpected leader transition is negative. In particular, they show that there is a 0.2 percentage point reduction in annual growth during the 5 years after an unexpected transition in leadership. Absent some disruptive effects, it is hard to see why the average effect of random leader transitions should be predictability negative.

RIFLE also offers practical advantages in terms of its generalizability. The Jones and Olken method requires not only identifying leader transitions due to death in office but also knowledge of the cause of the leader's death. While uncovering such information may be feasible for world leaders, it may be impractical or impossible in the case of more obscure leaders serving in less prominent offices. Even for the large U.S. cities that we study in this paper, we sometimes had difficulty finding the names of the mayors who served in the past. We suspect that finding detailed information about their cause of death would require heroic effort. Fortunately, our method does not require additional information beyond the leaders' terms of service, and so should be more applicable to studying a wide range of leaders.

We also note some limitations of our method, which are shared with others in the literature. Importantly, we cannot say anything about the performance of any individual leader or make direct comparisons of two leaders. We test a particular notion of what it means for leaders to matter, that variation in outcomes across leaders is greater than would be expected by chance. Our method enables us to assess whether the variation due to leaders is statistically significant in this sense, but not whether any individual leader's effect is significant.

We can think of our approach as testing the sharp null hypothesis that all leaders are equal, with respect to a particular outcome of interest. We conclude that leaders matter when we can reject the hypothesis that all leaders are equal. One reason we might fail to reject the null is if

selection reduces the variation among those who become leaders. For instance, electoral competition may result in only leaders above a certain level of quality being elected, or only leaders who share some common set of preferences and beliefs. If these leaders perform equally well, there will be no leader effects according to our definition. But such a null finding would not imply that replacing a sitting leader with a random person from the population would have no effects.

Some empirical studies in political economy ask about the effects of electing one type of leader or another. For example, in the context of U.S. mayors, there are studies on the effect of Democrats vs. Republicans (Ferreira and Gyourko 2009; Gerber and Hopkins 2011; Benedictis-Kessner and Warshaw 2016), females vs. males (Ferreira and Gyourko 2014), and those with vs. without business experience (Kirkland 2017). These kinds of studies can, in principle, demonstrate that leaders matter and they can further explore which dimensions of leaders' characteristics or backgrounds matter most for particular outcomes. However, this literature has produced a lot of null results, and even when results are not null, we might worry that many hypotheses were tested across many characteristics of leaders and many outcomes, leading us to wonder whether and how much variation in leaders really matters for a particular outcome of interest. In these cases, a more general test of leader effects is warranted. We see our method as a complement to (but not a replacement for) design-based studies of the effects of particular kinds of leaders. Iteration and interplay between our general test of leader effects and more specific testing of particular dimensions of leadership could be particularly fruitful in settings where leader effects have previously been elusive.

Monte Carlo Simulations Assessing the Properties of RIFLE

To assess the properties of RIFLE, we have conducted a series of Monte Carlo simulations. First, we show that if there are no leader effects, we will not detect them. Under the null hypothesis, p-values should be uniformly distributed between 0 and 1, and this is exactly what we find. We simulate data sets with a certain number of units and time periods. For our initial simulations, we assume that each leader's tenure is randomly drawn uniformly from integers between 1 and 5. We simulate an extra 5 periods for each unit and remove the first 5 periods. This ensures that the simulated data set starts in the middle of some leaders' tenures, making the simulations more similar to our subsequent analyses with real data. The outcome in the first period is drawn independently from a standard normal distribution, and the outcome in each subsequent period is a weighted average of growth in the previous period and a new draw from a standard normal distribution. This means there is random variation in growth from year to year and there is also serial correlation over time within each unit, but growth is unrelated to leaders.

In Figure 1, we present results from simulations with 20 units and 20 periods per unit, although results are similar even when we have as few as 5 units and 5 periods. We vary the extent of serial correlation in the data by varying the weight of the previous period's outcome in the current period's outcome, and we show results for weights of 0, i.e., no serial correlation, and .2, modest serial correlation. For each level of serial correlation, we simulate 1,000 data sets and implement RIFLE. Furthermore, to compare RIFLE to a more standard method, we also implement an F-test of joint significance after running a regression of the outcome on leader fixed effects.

The top row of Figure 1 shows the results of the F-tests, and the bottom row shows the results of RIFLE. When there is no serial correlation, both RIFLE and the F-test perform well, producing a uniform distribution of p-values as we would hope for when there are no leader effects. When we introduce serial correlation, however, RIFLE continues to perform well, while the F-test dramatically over-rejects the null. These simulations confirm that random noise and serial correlation do not contaminate the results of RIFLE as they do for more standard methods. Furthermore, our p-values are reliable even with a small number of units and periods. Also, notice that our procedure never requires the researcher to specify the nature of serial correlation in the data. If growth is unrelated to leaders' tenures, our test will not wrongly detect leader effects.

What if all leaders are equally able but there is a transition cost associated with political turnover? This is an important question because transition costs are a key concern for the methodology of Jones and Olken (2005). Because they focus on periods right around a political transition, such transition costs could lead them to significantly overstate leader effects. To assess the effect of transition costs for our method and that of Jones and Olken, we implement another battery of Monte Carlos, shown in the Appendix. As expected, transition costs strongly bias the Jones and Olken test against the null, while the implications of transition costs for RIFLE are much smaller. For transition costs to meaningfully bias our approach, these costs have to be very large, and in these cases, the bias is in the opposite direction, meaning that we under-reject the null. While transition costs would produce a false positive with the methodology of Jones and Olken, they lead our test to be conservative. The intuition is that every leader experiences exactly one transition cost during their tenure, so the estimated leader coefficients end up being similar in the real data and the r-squared is low. When we randomly permute the

leader tenures but keep the outcome data the same, some leaders end up with multiple transition costs in their tenure and others have none, which spreads out the estimated leader coefficients and increases the r-squared. However, again, for reasonably sized transition costs, the implications for our test are minimal.

What if leaders do not matter for a particular outcome but their terms of service are influenced by that outcome, perhaps because voters believe leaders matter and remove incumbents in bad times? In general, when leaders' terms of service are influenced by the outcome of interest, this poses a problem for our test, and the bias can go in either direction. We illustrate this problem with a theoretical model in the Appendix, and we supplement this with Monte Carlo simulations of the false discovery rate. Fortunately, when we conduct simulations using realistic parameter values generated from our data sets on real leaders, the extent of bias is negligible.

We've seen that RIFLE performs reasonably well when there are no leader effects. But if there are genuine leader effects, will it reliably detect them? The value of our method will be limited if there is not sufficient statistical power to detect meaningful effects when they do exist. Without having any specific hypotheses about which leaders are better, we don't know how we could achieve greater statistical power. As an example, our test should provide more power than that of Jones and Olken (2005) because instead of using only data from years around exogenous transitions, we're using all years and all leaders where data is available. Furthermore, the r-squared is an efficient statistic for our purposes because it implicitly puts more weight on the leader coefficients that are more precisely estimated, i.e., where there are more periods per leader. We have also explored alternative statistics like the standard deviation of the estimated leader coefficients, but these statistics are less efficient than the r-squared because many of those

coefficients are imprecisely estimated. In other words, we have designed our test with the goal of obtaining reliable p-values while also maximizing statistical power.

In Figure 2, we show the results of Monte Carlo simulations designed to explicitly assess the statistical power of our approach in different data sets. For each of the data sets that we subsequently analyze, we replace the actual outcomes with simulated data with known leader effects. Specifically, for each simulated data set, we generate a random variable from a standard normal distribution for each leader which indicates their individual effect. Then, for each observation in the data set, the outcome is simulated as random noise, drawn from a standard normal distribution, plus the leader effect which is multiplied by a number which corresponds to a particular magnitude of leader effects. For instance, when the leader effects are multiplied by one-ninth, this means that leader effects explain one-tenth of the variation in the outcome variable. For each simulated data set, we generate 19 permutations—corresponding to 20 different possible p-values, and we record whether the p-value is less than .05. Then, for each magnitude, we repeat this procedure 100 times and record how often we reject the null. Figure 2 plots the results of these simulations for all settings and for magnitudes of 0, .05, .1, .15, .2, .3, and .35.

As expected, when there are no leader effects, our test rejects the null about 5 percent of the time, as it should. When leader effects are small, say 5 to 10 percent of the variation in the outcome, our test will not reliably detect them in these settings. However, once leader effects increase to 15 to 25 percent, our test reliably detects them in most settings. And if leader effects explain 35 percent of the variation in the outcome or more, our test is virtually guaranteed to reject the null in all of our data sets. As expected, our power varies across data sets depending on the amount of data available. We have the most statistical power for our data set of world

leaders, including 153 countries over 135 years. In that case, we have a 70 percent chance of rejecting when leader effects explain 15 percent of the variation, and we're virtually guaranteed to reject if leader effects explain 20 percent. We have the least statistical power when we analyze the 31 countries that were classified as autocracies throughout their history and when we analyze U.S. mayors for the short period of 1986-2012. In those settings, we would need a 25 percent effect to have more than a 75 percent chance of rejecting the null. Therefore, our test is ill-equipped to detect small leader effects, but so long as leader effects explain more than 15 or 20 percent of the variation in the outcome of interest, we should reliably detect them in most settings.

If our test does lead us to reject the null, we have evidence that leaders matter, but that alone tells us nothing about the substantive size of leader effects. For many questions in the social sciences, including this one, our prior beliefs put little mass on the null hypothesis that an effect is exactly zero, and we care more about substantive significance than statistical significance (e.g., Ziliak and McCloskey 2008). As we've discussed, the r-squared statistic itself is not substantively interpretable. It tells us the proportion of variation in the data explained by leader fixed effects, but that number includes leader effects plus unit-specific time trends and the overfitting of random noise. However, the difference between the real r-squared and the average of the permuted r-squared statistics should increase with leader effects, so that difference should be useful for interpreting effect sizes. Even that number, however, does not have a direct substantive interpretation. One reason is that genuine leader effects should increase both the real r-squared and the permuted r-squared, although they should increase the former more than the latter. Therefore, if r-squared were greater in the real data than in the permuted data and one interpreted that difference as the proportion of variation attributable to leader effects, they would

understate the true effect. Therefore, in order to say more about effect sizes, we recorded this difference from the power simulations above.

Figure 3 shows the average difference between the real and permuted r-squared statistics for each of the settings in our subsequent analyses and for different magnitudes of leader effects ranging from 0 to 30 percent. As expected, there is not a one-to-one relationship between the difference in r-squared and the actual proportion of variation attributable to leader effects. Instead, the relationship is nonlinear, and as expected, the average slope is less than 1. The color coding is the same as in Figure 2, and we see that the slope tends to be lower when statistical power is lower. In any case, these simulation results, in conjunction with our actual results using real data, can be used to generate maximum likelihood estimates of effect sizes. For example, suppose that when we analyze the effects of U.S. mayors on employment, examining data from 1970 to 2015, we obtain an r-squared that is .01 greater than the average r-squared from the random permutations, that would imply that mayor effects explain more than 15 percent of the within-city variation in employment. For our subsequent results that do imply leader effects, we will utilize this approach to interpret the substantive size of those effects.

Applications to World Leaders, Governors, and Mayors

We examine the effects of three types of leaders on economic growth: national leaders; U.S. governors; and mayors of the top 100 U.S. cities. For mayors and governors, we also estimate leader effects on public finance decisions and crime rates. We utilize data on world leaders from Archigos (Goemans, Gleditch, and Chiozza 2009) version 4.1, and we collected data on U.S. governors and U.S. mayors from various sources. Our data sets indicate which leader held each leadership position in each year. In cases where multiple leaders served in the

same position in the same year, we record which leader held the position for the greatest portion of that year. We utilize data on GDP by country and year from the Maddison Project (Bolt and van Zanden 2014). Data on U.S. state- and county-level income per capita and employment come from the Regional Economic Information System (REIS) of Bureau of Economic Analysis. For U.S. cities, we focus on the 100 largest cities by 2015 population. Because annual income and employment estimates are not available by city, we match each city to its home county use county-level income and employment data from REIS. We drop the two cities that are not the largest in their county (Long Beach, CA and Mesa, AZ). Data on state- and city-level per capita property crime and violent crime come from the Federal Bureau of Investigation's Uniform Crime Reporting Program, and public finance data on state- and city- level revenues and expenditures comes from the U.S. Census Bureau's Census of Governments and Annual Surveys of Local Government Finances.

World Leaders

Following the methodology described above, we analyze the effects of world leaders by regressing GDP growth, after de-meaning the data by year, on a set of leader dummies. As shown in Table 1, the r-squared from this regression is .275 and the average r-squared from the permuted regressions is .246. The estimated p-value from 1,000 different permutations is .006. The difference between the real r-squared the average permutation of .030 implies that about 25 percent of the variation in economic growth within countries is attributable to variation in national leaders. In our data set, the standard deviation of growth, demeaned by country and year, is 6.0 percent. Therefore, as a country switches from an average leader to one that is one standard deviation above the mean, they can expect GDP growth to be about 1.5 percent higher

than normal. This is a substantively meaning number, and remarkably, it is nearly identical to the substantive size of leader effects implied by the results of Jones and Olken (2005, p. 837).

We next estimate leader effects separately for democracies and autocracies, as classified by the Polity IV scores. Jones and Olken (2005) classify countries according to their status at the time of a leader transition. Easterly and Pennings (2016), on the other hand, classify countries according to their average scores over time. Relying on average scores, however, means that countries will have a constant classification in the analysis even if they transition from autocracy to democracy, or vice versa, during the study period. To overcome this problem, we divide countries into three categories—those that were always democracies, those that were always autocracies, and those that changed their status at least once over the period of study. 29 countries in our data were always autocracies, 35 were always democracies, and 89 transitioned at some point. We implement our method separately for each subset.

We find some evidence of leader effects in autocracies ($p = .061$), strong evidence of leader effects in transitional countries ($p = .006$), and little evidence in democracies ($p = .463$). We should not conclude that leaders matter more in transitional countries than autocracies since we have more statistical power in the latter sample. The implied substantive sizes of leader effects are remarkably similar across both samples. For autocracies, the difference of .026 suggests that leader effects explain more than 25 percent of variation in growth (see Figure 3), and since the standard deviation of within-country growth in autocarcies is 7.4 percent, this implies that a leader who is one standard deviation above the mean will increase GDP growth by about 2 percentage points more than an average leader. For transitional countries, the difference of .034 suggests that leader effects explain 25 percent of the variation in economic growth, and since the standard deviation of within-country growth in transitional countries is 5.9 percent, this

implies that a leader who is one standard deviation above the mean will increase GDP growth by about 1.5 percent. Like Easterly and Pennings (2016) but unlike Jones and Olken's (2005), we find that leader effects are not limited to autocratic nations, but the implied effect sizes are slightly larger for autocracies than transitional countries. Like Jones and Olken but unlike Easterly and Pennings, we find little evidence for leader effects in democracies.

U.S. Governors

We extend our analysis to governors following the same basic methodology. Our state-level economic data starts in 1930 for personal income and in 1970 for employment. While these panels are shorter than those for world leaders, our power simulations suggest that we can still detect meaningful leader effects if they are present. As Table 2 shows, however, there is not much evidence of governor effects on these outcomes.

Although we find no evidence that governors matter for aggregate economic outcomes, perhaps governors differ in their use of various policy levers. To test for this, we examine state revenue excluding federal aid, state expenditures, and aid from the federal government to states from 1951-2008. Perhaps some governors are better than others at securing federal aid, and perhaps they raise and spend different amounts of money. We obtain positive point estimates for all three of these outcomes, and these estimates are statistically significant in the case of expenditures and federal aid ($p = .034$ and .018, respectively). The estimates imply that between 20 and 25 percent of variation in these public finance outcomes can be explained by the abilities and priorities of individual governors. Interestingly, although governors differ in their abilities or appetites to raise and spend money, those differences don't appear to translate into differences in state income and employment.

Perhaps the primary effects of governors are outside the economy, in which case we should shift our focus to outcomes that are potentially shaped more directly through state-level actions. Pursuing this idea, we next examine the effects of governors on crime rates. Using data from 1961 to 2012, we analyze the effects of governors on property and violent crime rates. We find statistically significant governor effects on property crime rates ($p = .044$), and suggestive effects for violent crime ($p = .120$). The differences between the real r-squared and the permuted r-squared, interpreted in conjunction with Figure 3, imply that governor effects explain 15-20 percent of the variation in both property crime and violent crime.

To the extent that we detect governor effects for public finance and crime, are these effect simply explained by differences between Democrats and Republicans? We address this question by removing the average effects of party before implementing our procedure. Specifically, instead of simply demeaning by time, we regress our outcomes on state fixed effects, year fixed effects, and an indicator for the governor's party, and we compute the residuals. When we do this, the estimated effects of governors on public finance and crime and virtually unchanged, suggesting that these effects are not explained by average differences between Democrats and Republicans. In fact, we estimate little average effect of party on these outcomes, suggesting that there must be meaningful variation in governors' abilities and priorities within a given party and state. This exercise demonstrates the flexibility of our method and its ability to address additional questions and disentangle potential mechanisms driving leader effects.

Mayors of the Top 100 U.S. Cities

We next study the effects of mayors for similar outcomes. We focus on the 100 largest U.S. cities according to 2015 population. For each city, we collected data on the names and

service dates of mayors dating back at least to 1970, when our data on local income and employment begin. We have obtained complete data for 70 of the top 100 cities. An important caveat is that aggregate, annual economic data for U.S. cities are not available. Instead, we use county-level economic estimates from REIS. As discussed above, we match each city to its county and drop the two cities in our data set that are not the largest city in their county. We can also weight the regressions according to the city's share of county population, which is allowable within our methodology and has no impact on our subsequent results.

Can mayors reasonably be expected to affect economic growth in their cities? The existing literature offers differing perspectives. One school of thought argues that Tiebout (1956) competition forces mayors to single-mindedly pursue economic development (Peterson 1981), with some going so far as to argue that cities are governed as "growth machines" that pursue development at the expense of all other priorities (Logan and Molotch 1987). Others contend that mayors are relatively weak executives that lack the power to control basic service delivery, much less to drive economic growth in their cities (Yates 1977). We find little evidence of mayoral effects on income and employment in their counties, as shown in Table 3. In both cases, mayoral fixed effects appear to explain a substantial amount of the variation in the outcome, but we see that the placebo dummies explain just as much, on average. Similarly, when we examine public finance data on city employees and the average salary of city employees, we continue to find null results.

We next turn to an analysis of mayoral effects on crime, an outcome over which we might expect mayors to have greater influence. After all, mayors directly appoint police chiefs and shape law enforcement policy within their jurisdictions. Nevertheless, we find no evidence that mayors affect either property or violent crime rates in their cities. Our results reveal no

evidence of mayoral effects for some of the most important outcomes in a city—the economy, the size of city government, and crime rates. Such results are generally consistent with the argument that mayors simply lack control over governance and service provision within their jurisdictions (e.g., Yates 1977). Furthermore, we continue to obtain similarly null results when we focus exclusively on the cities that, according to surveys conducted by the International City Management Association, have a mayor-council system in which the mayor has more independent authority. Our null results in this case could be the result of insufficient data, but we fail to detect mayor effects even for the outcomes and cities where we would most expect them.

Discussion and Conclusion

We have proposed a new approach for estimating leader effects—RIFLE. The primary innovation relies on random permutations for inference. Relative to methods previously used in the literature, ours makes use of more variation in the data and relies on fewer and weaker assumptions. In Monte Carlo simulations, RIFLE performs well even in relatively small samples. In addition, our method does not require knowledge of the particular circumstances surrounding leader transitions or the cause of leader deaths in office, making our approach easier to generalize to subnational or other less prominent offices where leader biographies are unlikely to be available.

Applying RIFLE in various political settings, we show that world leaders matter for economic growth, consistent with prior research in the same tradition. We also apply our method to settings where leader effects had not previously been estimated. We find no evidence that U.S. governors and mayors affect income and employment in their jurisdictions. We do find that governors, but not mayors, influence public finance in the forms of expenditure and federal aid.

And we also find some evidence that governors, but not mayors, affect crime in their jurisdictions. While previous studies have focused exclusively on aggregate economic outcomes, our results highlight the importance of matching different offices to relevant outcomes when estimating leader effects. We hope the application of our method to different contexts will improve our general understanding of where, when, and why leaders matter.

References

Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program. *Journal of the American Statistical Association* 105(490):493-505.

Acemoglu, Daron, Simon Johnson, and James A. Robinson. 2005. Institutions as the Fundamental Cause of Long-Run Growth. *Handbook of Economic Growth*, Philippe Aghion, and Steven N. Durlauf, eds., Amsterdam.

Benedictis-Kessner, Justin and Christopher Warshaw. 2016. Mayoral Partisanship and Municipal Fiscal Policy. *Journal of Politics* 78(4):1124-1138.

Bertrand, Marianne, and Antoinette Schoar. 2003. Managing with Style: The Effect of Managers on Firm Policies. *Quarterly Journal of Economics* 118(4):1169-1208.

Blinder, Alan S. and Mark W. Watson. 2016. Presidents and the U.S. Economy: An Econometric Exploration. *American Economic Review* 106(4):1015-1045.

Timothy Besley, Jose G. Montalvo and Marta Reynal-Querol. 2011. Do Educated Leaders Matter? *The Economic Journal* 121(554):F205-227.

Bolt, Jutta and Jan Luiten van Zanden. 2014. The Maddison Project: Collaborative Research on Historical National Accounts. *Economic History Review* 67(3):627-651.

Canes-Wrone, Brandice. 2006. *Who Leads Whom? Presidents, Policy, and the Public*. Chicago University Press, Chicago.

Carlyle, Thomas. 1859. *On Heroes, Hero Worship and the Heroic in History.* Wiley and Halsted, New York.

Corman, Hope, and Naci Mocan. 2005. Carrots, Sticks, and Broken Windows. *Journal of Law and Economics* 48(1):235-266.

Dewan, Torun, and David Myatt. 2008. The Qualities of Leadership: Communication, Direction and Obfuscation. *American Political Science Review* 102(3):351–368.

Dewan, Torun, and Francesco Squintani. 2015. On Good Leaders and Their Associates. Working paper <bit.ly/2kbHxn3>.

Easterly, William, and Steven Pennings. 2016. Shrinking dictators: how much economic growth can we attribute to national leaders? Working paper <bit.ly/2lz8TAs>.

Fisher, Ronald A. 1935. *Design of Experiments*. Oliver/Boyd, Edinburgh.

Ferreira, Fernando and Joseph Gyourko. 2009. Do Political Parties Matter? Evidence from U.S. Cities. *Quarterly Journal of Economics* 124(1):399-422.

Ferreira, Fernando and Joseph Gyourko. 2014. Does Gender Matter for Political Leadership? The Case of U.S. Mayors. *Journal of Public Economics* 112(April):24-39.

Gerber, Elisabeth R. and Daniel J. Hopkins. 2011. When Mayors Matter: Estimating the Impact of Mayoral Partisanship on City Policy. *American Journal of Political Science* 55(2):326-339.

Glaeser, Edward L., Rafael La Porta, Florencio Lopez-de-Silanes, and Andrei Shleifer. 2004. Do Institutions Cause Growth? *Journal of Economic Growth* 9(3):271-303.

Goemans, H.E., Kristian Skrede Gleditch, and Giacomo Chiozza. 2009. Introducing *Archigos*: A Data Set of Political Leaders, 1875-2003. *Journal of Peace Research* 46(2):269-283.

Ho, Daniel E. and Kosuke Imai. 2006. Randomization Inference with Natural Experiments: An Analysis of Ballot Effects in the 2003 California Recall Election. *Journal of the American Statistical Association* 101(475):888-900.

Jones, Benjamin F., and Benjamin A. Olken, (2005) Do Leaders Matter? National Leadership and Growth since World War II. *Quarterly Journal of Economics* 120(3):835-864.

Kennedy, Sean, and Parker Abt. 2016. Trump is right about violent crime: It's on the rise in major cities. *The Washington Post*, August 5.

Kirkland, Patricia A. 2017. The Business of Being Mayor: Mayors and Fiscal Policy in U.S. Cities. Working paper <bit.ly/2wYKwS5>.

Levi, Margaret, and Josh Ahlquist. 2011. Leadership: What it means, what it does, and what we want to know about it. *Annual Review of Political Science* 14:1-24.

Logan, John, and Harvey Molotch. 1987. *Urban Fortunes: The Political Economy of Place*. University of California Press, Berkeley.

Marx, Karl. 1852. The Eighteenth Brumaire of Louis Napoleon. *Die Revolution*, New York.

North, Douglas. 1990. *Institutions, Institutional Change, and Economic Performance.* Cambridge University Press, Cambridge.

Peterson, Paul. 1981. *City Limits*. University of Chicago Press, Chicago.

Rosenbaum, Paul R. 2002. *Observational Studies*, 2nd edition. Springer, New York.

Sanburn, Josh. 2016. Chicago Is Responsible for Almost Half of the Increase in U.S. Homicides. *Time Magazine*, September 19.

Tiebout, Charles. 1956. A Pure Theory of Local Government Expenditures. *Journal of Political Economy* 64(5):416-424.

Wooldridge, Jeffrey M. 2010. *Econometric Analysis of Cross Section and Panel Data*. MIT Press, Cambridge.

Yao, Yang and Muyang Zhang. 2015. Subnational Leaders and Economic Growth: Evidence from Chinese Cities. *Journal of Economic Growth* 20(4):405-436.

Yates, Douglas. 1977. *The Ungovernable City: The Politics of Urban Problems and Policy Making*. MIT Press, Cambridge.

Ziliak, Stephen T. and Deirdre N. McCloskey. 2008. *The Cult of Statistical Significance: How the Standard Error Cost Us Jobs, Justice, and Lives*. University of Michigan Press, Ann Arbor.

Table 1. Results for World Leaders and GDP, 1876-2010

subset	r^2	avg permutation	difference	p-value
All Countries	.275	.246	.030	.006
Autocracies	.176	.150	.026	.061
Democracies	.389	.388	.001	.463
Transitional	.279	.240	.038	.004

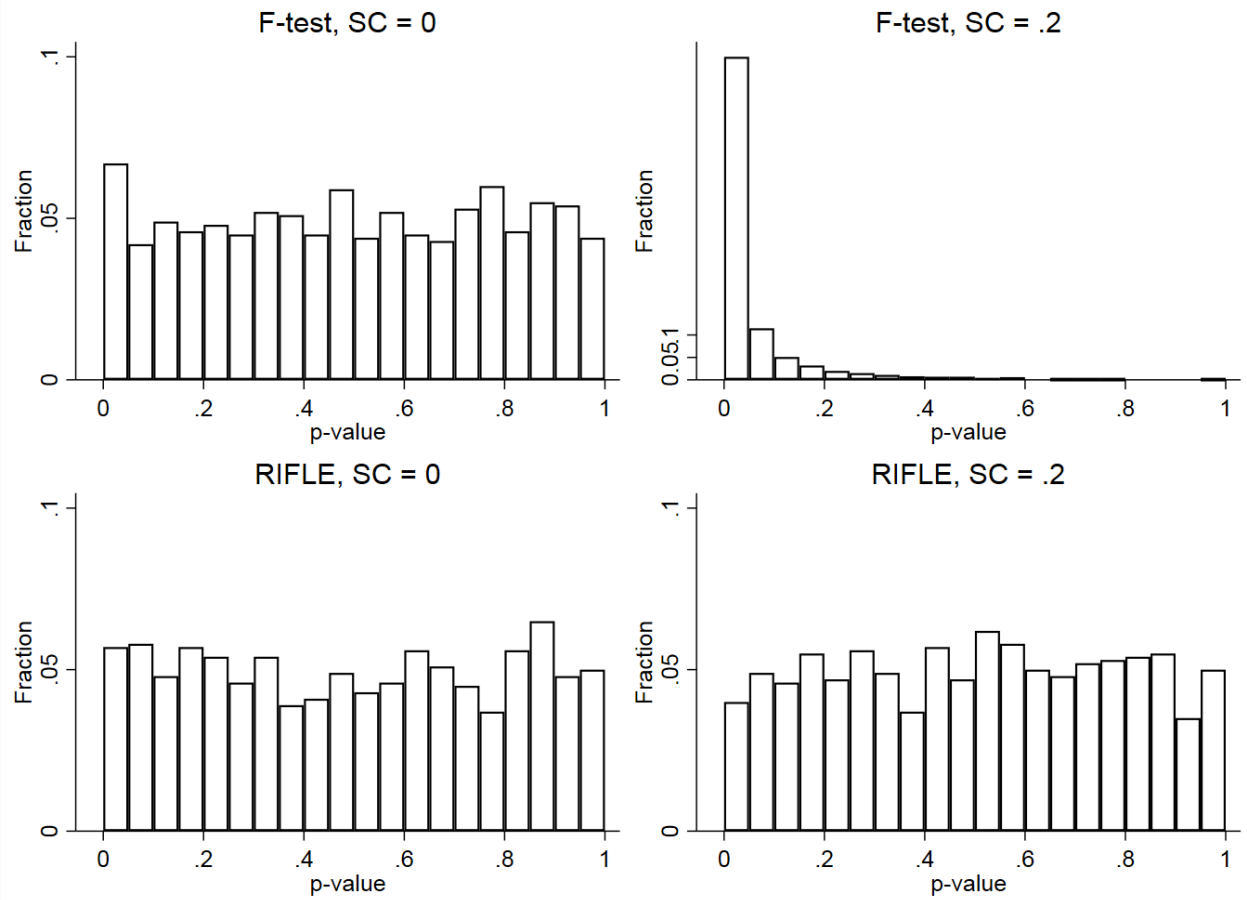
Table 2. Results for U.S. Governors

outcome	years	r^2	avg permutation	difference	p-value
Income	1930-2015	.166	.154	.013	.219
Employment	1970-2015	.450	.500	−.051	.994
Revenue	1951-2008	.230	.204	.025	.207
Expenditures	1951-2008	.203	.180	.023	.034
Federal Aid	1951-2008	.135	.116	.018	.018
Property Crime	1961-2012	.182	.167	.016	.044
Violent Crime	1961-2012	.173	.159	.014	.120

Table 3. Results for U.S. Mayors

outcome	years	r^2	avg permutation	difference	p-value
Income	1970-2015	.174	.183	−.009	.767
Employment	1970-2015	.608	.606	.002	.413
Public Employment	1970-2010	.183	.207	−.024	.667
Avg Public Salary	1973-2010	.098	.108	−.010	.799
Property Crime	1986-2012	.191	.198	−.007	.743
Violent Crime	1986-2012	.188	.193	−.005	.683

Figure 1. Monte Carlos with Random Noise, Serial Correlation, but No Leader Effects



Each histogram shows the distribution of p-values resulting from 1,000 simulated data sets with serial correlation (SC) of varying magnitudes. The top row presents results from a standard F-test, and the bottom row presents tests using RIFLE. See the text for more details.

Figure 2. Statistical Power across Settings

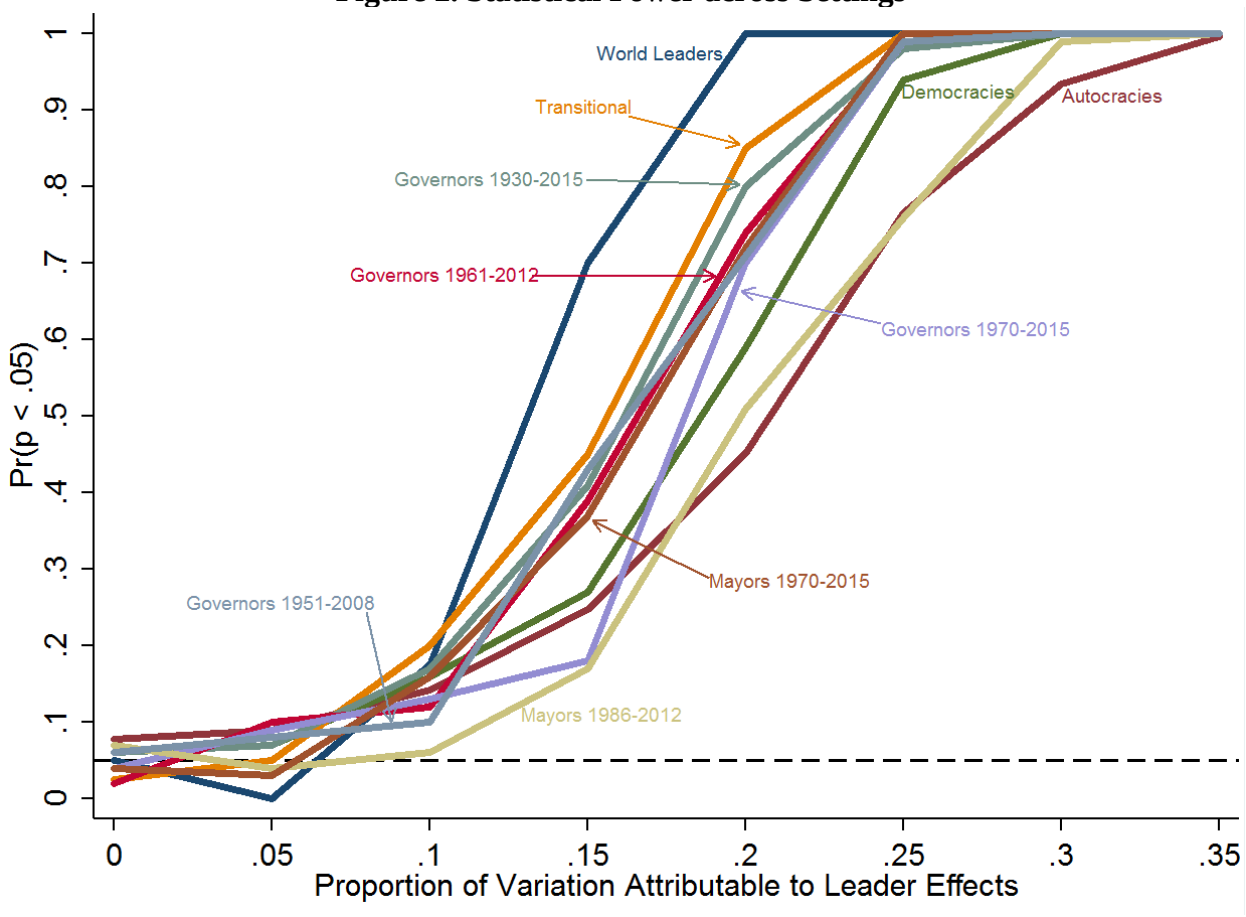
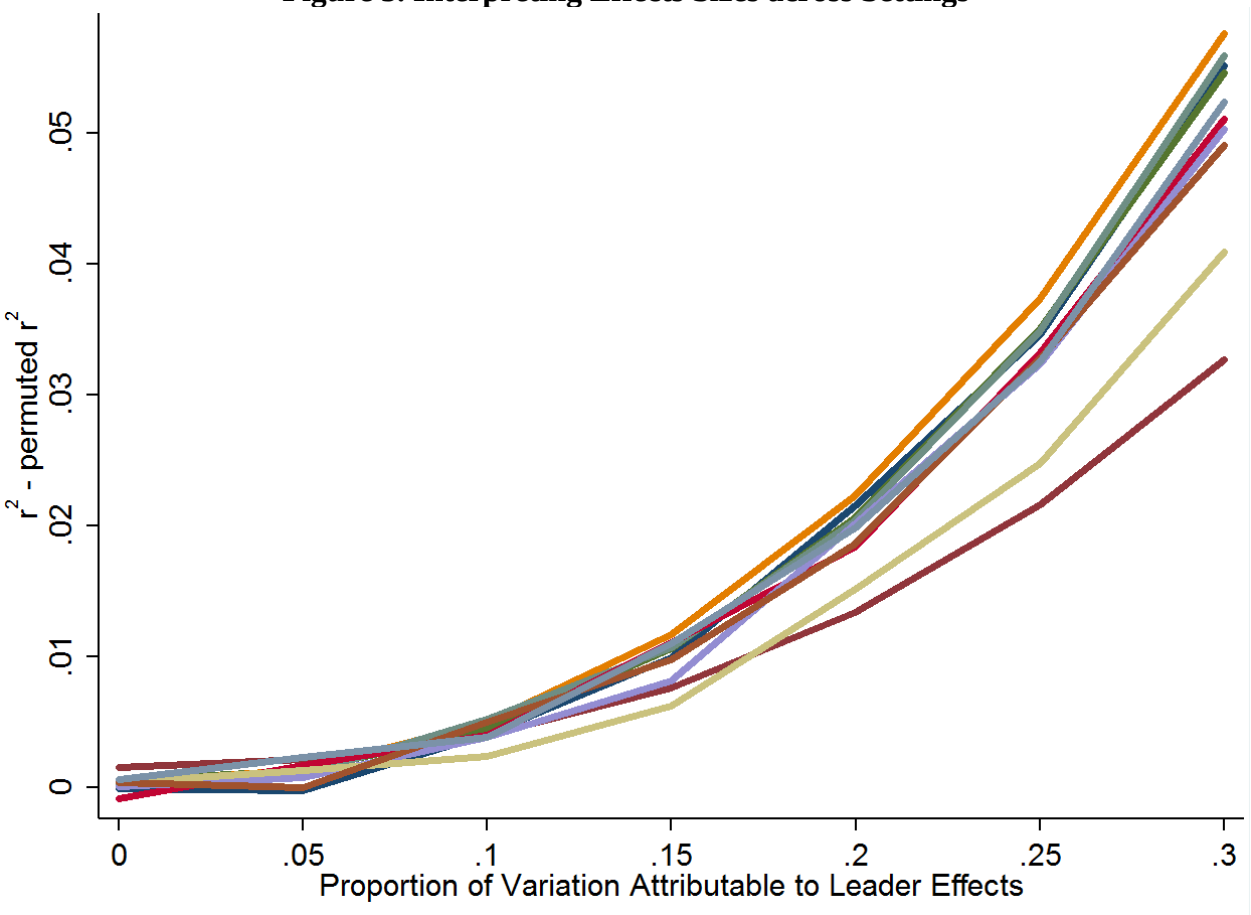


Figure 3. Interpreting Effects Sizes across Settings



The color coding for each setting is the same as in Figure 2.

Appendix

Transition Costs

To explore the implications of transition costs for our test, we conducted an additional battery of Monte Carlo simulations. For simplicity, everything is identical to the analyses in Figure 1 with 20 units and 20 periods except there is no serial correlation and there is a transition cost in the first period each leader takes office. Specifically, the outcome in each period is drawn from a normal distribution with a mean of 0 and standard deviation of 1, and a constant amount is subtracted in transition periods. This means that in non-transition years, the mean and standard deviation are 0 and 1, respectively, while in transition years the mean and standard deviation are −X and 1, respectively, where X corresponds to the magnitude of the transition cost. In Figure A1, we show results for transition costs of .1, .25, .5, 1, and 2.

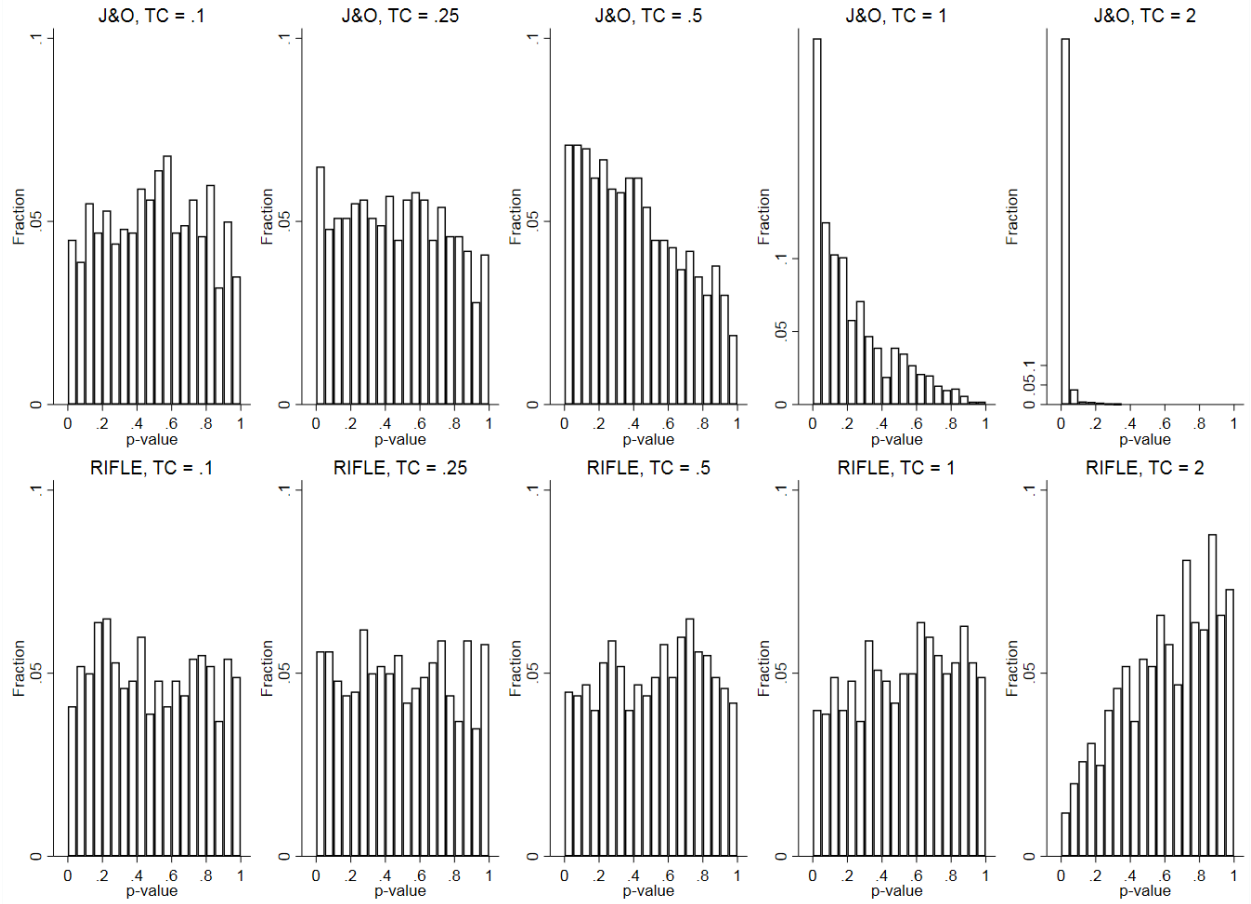
For each transition cost, we simulate 1,000 data sets and implement RIFLE along with a test in the spirit of Jones and Olken (2005). When implementing the latter test, we compute the absolute change in growth for each period, i.e., $|Y_t - Y_{t-1}|$, and we test whether this absolute change is greater in transition versus non-transition years using a t-test.

Figure A1 shows the distribution of p-values resulting from both tests across different transition costs. As expected, the Jones and Olken test (top row of Figure A1) produces p-values that are skewed right, meaning that this test over-rejects the null hypothesis. Furthermore, the transition cost need not be too large to generate a significant bias. On the other hand, RIFLE performs much better in the presence of transition costs. Unless the transition cost is larger than the standard deviation of the outcome in non-transition years, the bias introduced by transition costs is negligible. Furthermore, when a bias is detectable, it goes in the opposite direction.

Specifically, the distribution of p-values is skewed to the left, meaning that RIFLE under-rejects the null. The intuition for this result is provided in the main text.

The test of Jones and Olken overstates leader effects in the presence of transition costs, while RIFLE performs much better. When there is a meaningful bias for RIFLE, which only occurs for very large transition costs, it leads us to understate leader effects. Overall, the implications of transition costs are minimal for RIFLE, and if anything, they lead RIFLE to be a conservative test of leader effects.

Figure A1. Effect of Transition Costs for Jones and Olken vs. RIFLE



Each histogram shows the distribution of p-values resulting from 1,000 simulated data sets with transition costs (TC) of varying magnitudes. The top row presents results from a test in the spirit of Jones and Olken (2005), and the bottom row presents tests using RIFLE. See the text for more details.

Retrospective Voting

As discussed in the main text, one concern for our test is that the outcome of interest influences the tenures of leaders. Suppose, for example, that governors do not matter for a particular outcome, but voters believe that they do, and therefore, the values of that outcome variable influence the chances that the governor will stay in office. This kind of process, which for convenience we'll refer to as *retrospective voting*, could potentially bias the results of RIFLE. As we'll show with a theoretical model, the bias could go in either direction, leading us to either over- or under-reject the null. Fortunately, as we'll show with Monte Carlo simulations, the extent of this bias is negligible for realistic levels of retrospective voting in our substantive settings.

Theoretical Model

To understand and illustrate the bias that arises from retrospective voting, consider the following theoretical model. Suppose there is a binary outcome of interest (i.e., welfare in each term is either good or bad), all leaders are the same (i.e., the outcome is unrelated to the identity of the leader), there is a two-term limit, and retention for first-termers depends entirely on the outcome in the first term.

Recall that $r^2 \equiv 1 - \frac{RSS}{TSS}$, where RSS is the residual sum of squares and TSS is the total sum of squares. With RIFLE, the TSS is identical for both the real data and the permuted data sets where the ordering of leaders is randomly shuffled. Therefore, to think about how RIFLE will perform, we can focus on the RSS. Under the null, we'd like the RSS to be identical, in expectation, for the real data and the permuted data. If the expected RSS is greater in the real data, that means the r-squared will be smaller, and we will under-reject the null. If the expected

RSS is smaller in the real data, the r-squared will be larger, and we will over-reject the null. Since the sample size is the same for both the real and permuted data sets, we can also think about the average squared residual, and by comparing the expected average squared residual in the real and the permuted data, we can assess whether RIFLE will over- or under-reject the null.

In this theoretical model, the data set of leaders and outcomes will include only three different kinds of leaders. Anyone who has a bad year in their first term will be removed from office, so there will be one-termers who had a bad outcome—let's refer to this type of leader as 0. To serve two terms, the outcome must have been good in the first term, but the outcome could have been either good or bad in the second term, so in addition to 0's there are also 1-0's and 1-1's.

Because leaders don't matter, the probability of a good outcome does not depend upon who is in office. Let's call the probability of a good term p, where $0 < p < 1$. In expectation, the proportion of 0's among all leaders is $1 - p$, the proportion of 1-0's is $p(1 - p)$, and the proportion of 1-1's is p^2. As a share of all observations (i.e., terms) in the data set, 0's comprise $\frac{1-p}{1+p}$ of them, 1-0's comprise $\frac{2p-2p^2}{1+p}$, and 1-1's comprise $\frac{2p^2}{1+p}$.

Let's calculate the expected average squared residual in the real data. The squared residual for the 0's and the 1-1's is 0. In both cases, their leader fixed effect perfectly fits every data point. For the 0-1's, the predicted values will be $\frac{1}{2}$, the residuals will be $\frac{1}{2}$, so the squared residuals will be $\frac{1}{4}$. Since 1-0's comprise $\frac{2p-p^2}{1+p}$ of all observations in our data set, the expected average squared residual in the real data set is $(\frac{1}{4})(\frac{2p-2p^2}{1+p}) = (\frac{1}{2})(\frac{p-p^2}{1+p})$.

In our random permutations, instead of three kinds of leaders, there will now be six: 0, 1, 0-0, 0-1, 1-0, and 1-1. Four of these types have no variation in their outcome so they make no

contribution to the RSS, whereas the 0-1's and the 1-0's will again have residuals equal to $\frac{1}{4}$. Those two types will comprise $2p^2(1-p)$ of the leaders and $\frac{4p^2(1-p)}{1+p}$ of the terms. Therefore, the expected average squared residual in the permutated data set is $\left(\frac{1}{4}\right)\left(\frac{4p^2(1-p)}{1+p}\right) = p(\frac{p-p^2}{1+p})$.

Comparing the expected average squared residuals in the real and permuted data, we see that they equal each other if and only if $p = \frac{1}{2}$. If $p > \frac{1}{2}$, the squared residuals will be greater in the permuted data, meaning the r-squared is lower, and RIFLE will over-reject the null. Alternatively, if $p < \frac{1}{2}$, the squared residuals will smaller in the permuted data, meaning the r-squared will be greater, and RIFLE will under-reject the null. In other words, a process like retrospective voting can produce a bias, and that bias can go in either direction.

To gain some intuition for the bias in the model, consider an extreme case where p is very close to zero but still some positive number. Remember that the r-squared of the regression of the outcome on leader fixed effects is determined entirely by the proportion of two-term leaders that have one good term and one bad term. In the rare case when there is a good outcome and a leader is retained, they will almost certainly be a 1-0. 1-1's will be exceedingly rare relative to 1-0's. This means that almost all of the two-termers in the real data will be 1-0's, contributing positively to the RSS. When we permute the leader tenures, most of those two-term leaders will happen fall on two bad terms, and they'll become 0-0's, where they will not add to the RSS. The r-squared will be very high in both cases, but it will be almost exactly 1 in the permuted data, and it will be slightly lower in the real data, meaning that RIFLE will under-reject the null.

Similarly, consider an extreme cases where p if very close to but still less than 1. Almost all leaders are 1-1's, and there are roughly equal shares of 0's and 1-0's. When this is a rare bad

outcome, half of those belong to one-termers, in which case they contribute nothing to the RSS. However, most of the leaders are two-termers, which means that in the random permutations, most of the bad outcomes get assigned to two-termers, creating 1-0's or 0-1's and increasing the RSS. Again the r-squared is very high in both cases, but it's higher in the real data than the permuted data, meaning that we over-reject the null.

Monte Carlos

To understand whether this bias is relevant for our substantive settings, we have implemented Monte Carlo simulations where the data generating process presumably matches the real world more closely than our illustrative model with binary outcomes, a two-term limit, and retention determined entirely based on the outcome. Specifically, we assume the outcome is normally distributed but with serial correlation as in Figure 1. We explore different levels of serial correlation by varying the weight of the previous outcome on the current outcome.

Leader retention is probabilistic, where the probability of retention depends on the outcome in the most recent period. Consistent with our data on world leaders in democracies, we set the average probability of retention in each period as .73, and we allow that probability to vary according to a particular degree of retrospective voting. For example, a retrospective voting value of .02, which is most consistent with our data on GDP in democracies, means that when the outcome is one standard deviation above average, the leader is two percentage points more likely to be retained.

Table A1 shows the results of these Monte Carlo simulations for five different values of serial correlation and seven different values of retrospective voting. For each set of values, we simulated 1,000 data sets with 20 units and 20 periods, and we implemented RIFLE. The table

reports the proportion of simulations in which RIFLE rejected the null. If there is no bias, this false discovery rate should be .05.

Consistent with our previous simulations, there is no bias when there is no retrospective voting. Interestingly, when the serial correlation is low, we see that retrospective voting leads us to under-reject the null in this setting, and when serial correlation is high, retrospective voting leads us to over-reject the null. Fortunately, for the parameter values that most closely match our substantive data sets (retrospective voting = .02 and serial correlation = .4), the bias is negligible. Therefore, although a process like retrospective voting can induce a bias, this bias is likely minimal for the substantive settings studies in this paper.

Table A1. False Discovery Rate under Serial Correlation and Retrospective Voting

		relationship between performance and retention rate						
		0	.02	.04	.08	.16	.32	.64
serial correlation	0	.045	.051	.042	.055	.033	.017	.006
	.2	.065	.050	.045	.055	.052	.070	.044
	.4	.054	.043	.031	.054	.057	.094	.155
	.6	.040	.056	.045	.056	.056	.098	.236
	.8	.054	.051	.046	.060	.057	.080	.181