On the Validity of the Regression Discontinuity Design for Estimating Electoral Effects: New Evidence from Over 40,000 Close Races

Andrew C. Eggers London School of Economics
Anthony Fowler University of Chicago
Jens Hainmueller Stanford University
Andrew B. Hall Harvard University
James M. Snyder, Jr. Harvard University and NBER

The regression discontinuity (RD) design is a valuable tool for identifying electoral effects, but this design is only effective when relevant actors do not have precise control over election results. Several recent papers contend that such precise control is possible in large elections, pointing out that the incumbent party is more likely to win very close elections in the United States House of Representatives in recent periods. In this article, we examine whether similar patterns occur in other electoral settings, including the U.S. House in other time periods, statewide, state legislative, and mayoral races in the U.S. and national or local elections in nine other countries. No other case exhibits this pattern. We also cast doubt on suggested explanations for incumbent success in close House races. We conclude that the assumptions behind the RD design are likely to be met in a wide variety of electoral settings and offer a set of best practices for RD researchers going forward.

In recent years, the regression discontinuity (RD) design has become widely used in political science. In general, RD designs are used to estimate the effect of a treatment that changes discontinuously at a threshold value of a continuous variable. In the first application, for example, Thistlethwaite and Campbell (1960) measured the effect of a scholarship by comparing the subsequent performance of students whose test scores were just high enough to win the scholarship to that of students who narrowly fell short.[1] In political applications, the most common use of RD has been to measure the effect of election results on various political and economic outcomes of interest.[2] These applications take advantage of the fact that in two-candidate plurality

Andrew C. Eggers is Assistant Professor, Department of Government, London School of Economics, Houghton Street, London WC2A 2AE (a.c.eggers@lse.ac.uk). Anthony Fowler is Assistant Professor, Harris School of Public Policy Studies, University of Chicago, 1155 E. 60th Street, Room 165, Chicago, IL 60637 (anthony.fowler@uchicago.edu). Jens Hainmueller is Associate Professor, Department of Political Science and Graduate School of Business, Stanford University, 616 Serra Street, Encina Hall West, Room 100, Stanford, CA 94305-6044 (jhain@stanford.edu). Andrew B. Hall is a Ph.D. Candidate, Department of Government, Harvard University, 1737 Cambridge St. K453, Cambridge, MA 02138 (hall@fas.harvard.edu). James M. Snyder, Jr. is Professor of Government, Harvard University, 1737 Cambridge St., Cambridge, MA 02138 (jsnyder@gov.harvard.edu). Snyder is also a Research Associate at the National Bureau of Economic Research.

For generously providing data, the authors thank Alberto Abadie, Melissa Dell, Fernando Ferreira, Alexander Fouirnaies, Ronny Freier, Danny Hidalgo, Yusaku Horiuchi, and Carl Klarner. For helpful comments, we thank Devin Caughey, Justin Grimmer, Gary King, and Jas Sekhon. We especially thank Olle Folke for his collaboration on earlier drafts of this article as well as his enthusiastic support throughout the project. The data used in this study can be downloaded for replication from the AJPS Data Archive on Dataverse (http://dvn.iq.harvard.edu/dvn/dv/ajps).

[1]The data used in this study can be downloaded for replication from the AJPS Data Archive on Dataverse. The Supporting Information (SI) is posted on the AJPS website.

[2]Examples include Lee, Moretti, and Butler (2004), DiNardo and Lee (2004), Hainmueller and Kern (2008), Leigh (2008), Pettersson-Lidbom (2008), Broockman (2009), Butler (2009), Dal Bó, Dal Bó, and Snyder (2009), Eggers and Hainmueller (2009), Ferreira and Gyourko (2009), Uppal (2009, 2010), Cellini, Ferreira, and Rothstein (2010), Ade and Freier (2011), Gerber and Hopkins (2011), Trounstine (2011), Boas and Hidalgo (2011), Folke and Snyder (2012), Gagliarducci and Paserman (2012), and Dell (2012).

American Journal of Political Science, Vol. 59, No. 1, January 2015, Pp. 259–274

 DOI: 10.1111/ajps.12127

elections, the "treatment"—winning the election—is applied to any candidate who surpasses the vote share threshold of 50%.[3] The discontinuous relationship between electoral success and political support makes the RD design an intuitively appealing strategy for estimating the effects of election outcomes: because the treatment depends only on a threshold value of political support, candidates or parties that receive just enough support to win may be fundamentally similar (and thus comparable) to candidates or parties that narrowly lose.

Three recent papers suggest that, despite the intuitive appeal of the RD design, the winners and losers of close elections may not in fact be comparable. Jason Snyder (2005) shows that in U.S. House elections between 1926 and 1992, incumbents won a disproportionate share of very close races. Caughey and Sekhon (2011) investigate this further and show, among other things, that winners in close U.S. House races raise and spend more campaign money. Grimmer et al. (2012) show that U.S. House candidates from the party in control of state offices, such as the governorship, secretary of state, or a majority in the state house or state senate, hold a systematic advantage in close elections.[4] Interpreted most narrowly, these studies suggest that the electoral RD design cannot be applied in a straightforward manner to U.S. House elections, given that the winners and losers of close races for this legislature appear to differ systematically. More broadly, these studies cast doubt on the enterprise of the electoral RD design, given that the manipulation necessary to produce such systematic differences would likely afflict close elections in other electoral settings as well.[5]

In this article, we consider the validity of electoral RDs from an empirical and theoretical perspective in light of these critiques. We review the assumptions behind the electoral RD design as formalized by Lee (2008) and consider their applicability to close elections. We then assess whether the evidence of systematic incumbent advantages in the U.S. House indicates a general problem with the use of RD to measure electoral effects. First, we assess whether similar problems arise in other electoral settings, including every partisan, single-winner, plurality/majoritarian election setting where data could be collected and assembled. We study elections to the U.S. House in other time periods as well as statewide, state legislative, and mayoral races in the United States; we also study national and/or local elections in the United Kingdom, Canada, Germany, France, Australia, New Zealand, India, Brazil, and Mexico.

We do not find a single other case that exhibits systematic incumbent advantages. We then consider from a theoretical perspective the mechanisms that could produce the type of incumbent advantages that have been detected in the post–World War II U.S. House, concluding that existing explanations are not convincing. This suggests that the unusual success of incumbents in very close House elections might result from chance rather than the ability of incumbent candidates to manipulate outcomes in this context and that evidence of incumbent dominance in close U.S. House elections does not pose a general threat to the validity of RD designs in electoral settings.

We conclude the article by providing recommendations to future researchers estimating electoral effects using RD designs. Consistent with Caughey and Sekhon (2011), we argue that the burden is on empirical researchers to justify their assumptions with theory and data. We advocate a three-step procedure combining theory and data analysis that should guide researchers in assessing the validity of an electoral RD in a particular setting. We pay particular attention to the problem of multiple testing, noting that statistical imbalance is expected to arise by chance from time to time and does not automatically invalidate the underlying assumption of an RD design, and we also point out that the RD design may continue to be the best available estimator even when imbalances are present, relying as it does on more transparent and plausible assumptions than available alternatives.

In short, despite recent concerns, we believe that the RD design is a fundamentally sound and widely applicable approach to learning about the effect of election results on a variety of political and economic outcomes. Although there are potentially many issues with applying RD designs to any particular setting, the evidence of incumbent dominance in very close U.S. House elections over the post-WWII period does not appear to uncover any fundamental problem with electoral RD designs.

[3]More generally, in any plurality election, a candidate's result is a discontinuous function of her vote share, with a threshold that depends on the performance of other candidates.

[4]We are also aware of one other working paper identifying a potential concern with the RD design in close elections. Vogl (2012) finds that black candidates are better at winning close races than their white opponents in mayoral races in the U.S. South (but not elsewhere). However, the statistical evidence is weak since there have been very few close mayoral races in the South between a white and black candidate. In Vogl's sample, there are only 38 such cases (from 18 unique cities) where the margin of victory was less than 20 points.

[5]Substantively, these studies also raise the prospect of fraud in close U.S. House races. Here, we focus on methodological implications, although we briefly discuss this issue later in the paper.

The Comparability of Winners and Losers of Close Elections

The intuitive appeal of the RD design in the analysis of elections derives from the idea that candidates who win and lose close elections should be comparable on average. This comparability depends on the assumption that the candidates or parties under consideration do not have complete control over the vote share they receive. If this were not the case (e.g., if better-resourced candidates could examine their opponent's final vote total and then decide whether to increase their own) then the winners and losers of close elections may well differ systematically. Lee (2008) formalizes this logic, showing that a comparison of narrow winners and losers identifies the average treatment effect of winning at the threshold as long as there is an exogenous random chance component to candidates' vote shares that has a continuous density (also see Hahn, Todd, and Van der Klaauw 2001).

A priori, the fundamental continuity assumption that implies candidates do not perfectly control the electoral outcome seems likely to be met, not just because the weather or far-off current events can influence outcomes (a common justification offered in electoral RD studies), but also because every close election involves (at least) two candidates; the fact that no candidate can control the campaign activities of her opponent would seem to be a strong indication that she cannot perfectly control her own vote share. Nevertheless, in principle it is, of course, possible that certain types of candidates could have a degree of precise control over electoral outcomes that would render the electoral RD design invalid. For example, if incumbent candidates had a systematic ability to convert narrow losses to narrow victories through some combination of legal challenges, electoral fraud, and heroic campaign feats, then close winners and losers would no longer be comparable and the RD design might no longer identify the effect of the electoral outcome.

As noted above, recent evidence suggests that winners and losers are not in fact comparable in close elections for the U.S. House of Representatives. Winners of close elections appear to be disproportionately incumbents (Snyder 2005); they also appear to be disproportionately aligned with the locally dominant party (Grimmer et al. 2012) and, among other things, have more experience and money (Caughey and Sekhon 2011). It is easy to see why such candidates would in general be more electorally successful, but it is less clear why they would disproportionately win what should be essentially coin flips, according to the theory laid out in Lee (2008).

FIGURE 1 Proportion of Previous Democratic Wins as Function of Democratic Vote Margin, U.S. House, 1946-2010

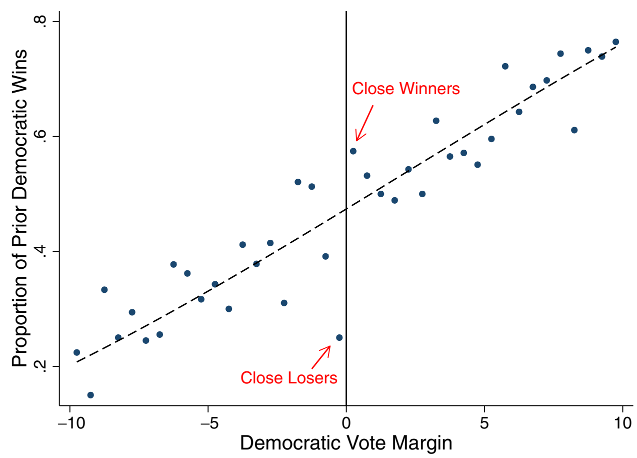


Figure 1 offers one view of the problem in the U.S. House of Representatives for the period from 1946 to 2010. For each 0.5 point bin of Democratic vote margin (e.g., all elections where the Democratic margin of victory was between 1.5 and 2 percentage points), we plot the proportion of cases in which a Democrat won the district in the *previous* election. As expected, there is a smooth, positive relationship between the Democratic margin of victory and the proportion of cases in which a Democrat was an incumbent. However, if we look at the bins immediately on either side of 0, we see a strange phenomenon. In the 59 total cases in which the Democrat *won* by less than half a percentage point (i.e., the first bin to the right of the threshold that is equivalent to Democratic vote percentages between 50 and 50.25), a Democrat previously won the seat almost 60% of the time; in the 54 total cases in which the Democrat *lost* by less than half a percentage point (i.e., the first bin to the left of the threshold that is equivalent to Democratic vote percentages between 49.75 and 50), a Democrat previously won the seat only 25% of the time. Within this sample of extremely close elections, we would expect the incumbent party to lose the seat just as often as it wins, but it appears to win a disproportionate share of close races. This highlights the exception first identified by Snyder (2005) and pursued further by Caughey and Sekhon (2011).

What accounts for the disproportionate success of the incumbent party in close U.S. House races? Snyder (2005) interprets it as evidence of corrupt electoral manipulation, suggesting that the complexity of the process of collecting and tabulating votes in close elections leaves opportunities for incumbent candidates to somehow tamper with the results of close elections. Grimmer et al. (2012) expand on these ideas in an analysis of a longer period of U.S. House races (1880–2008), showing that (particularly in the earlier period) candidates from the party that controlled local and state offices had a similarly substantial advantage; they suggest that part of the reason why "structurally advantaged candidates" disproportionately win close elections is that they are more successful in post-election legal battles. While conceding that a convincing explanation for this sorting remains elusive, Caughey and Sekhon (2011) point to the ability of well-organized campaigns to obtain precise information about likely outcomes and to take extraordinary measures to secure victory in very close races.

We return to these explanations for sorting in U.S. House elections below. For now, we note that the evidence of sorting in close U.S. House elections appears to cast doubt on the validity of RD as a strategy for measuring electoral effects not just in the U.S. House but also in a much broader class of electoral contexts. Although close U.S. House races are different in some respects from close races in most other settings (e.g., more money raised and spent, more polling conducted), there would seem to be at least as much scope for precise manipulation of outcomes in many other contexts. In legislative elections in many developing democracies, for example, electoral fraud is more common than in closely monitored U.S. House contests (Lehoucq 2003; Simpser 2013). Polling technology is less widely used in most settings where researchers are interested in using RD to measure electoral effects, but in many of these settings the electorate is much smaller, such that candidates arguably have similarly precise information about likely outcomes. The existing evidence of systematic incumbent advantages in close U.S. House elections may therefore pose a general threat to the validity of RD-based electoral studies.

In the subsequent sections, we assess the nature of this threat by examining evidence from other electoral settings. This evidence informs our subsequent theoretical analysis which asks what mechanisms could account for the anomalous patterns in the U.S. House.

Why Focus on Incumbency?

In principle, in electoral RD designs, as in other RD designs, one could check for differences between narrow winners and losers in as many pre-election characteristics as one can measure. In assessing the validity of electoral RD designs across various political settings, we focus on the role of incumbency: does the incumbent party disproportionately win close elections? We focus on incumbency for three reasons, which we can characterize roughly as an empirical reason, a statistical reason, and a theoretical reason.

The empirical reason for focusing on incumbency is that although existing studies have pointed out differences between winners and losers in a variety of characteristics, all of these differences can be viewed as proxies for incumbency. Caughey and Sekhon (2011) test for imbalances in the largest set of background covariates, showing that in addition to the incumbent party, candidates who received a higher vote share in the previous election, spent more money, or were predicted to win (among other differences), were more likely to win very close elections.[6] As shown by Table 1, however, the covariates Caughey and Sekhon (2011) study are so highly correlated with the party of the incumbent that after controlling for the party of the incumbent, the evidence of imbalance in the other covariates disappears. In the leftmost column of that table, we report the full list of covariates for which Caughey and Sekhon (2011) find substantial imbalance. To document imbalance, they restrict attention to close elections (defined as those with a margin of less than half a percentage point) and compute the mean difference for each covariate between districts in which the Democrat wins and districts where the Democrat loses. The middle column (labeled "Original Specification") reports the p-value corresponding to their test of the null hypothesis that this expected difference is zero.[7] In the rightmost column, we report p-values from another analysis that differs only in that incumbency (i.e., "Democratic Win") is added as a control.[8] The fact that none of these p-values is below .1 indicates the high degree of collinearity

[6]Caughey and Sekhon (2011) report that barely winners received more campaign contributions and spent significantly more money than barely losers. In testing for these imbalances, they are careful to use a measure of contributions that removes those made *after* Election Day. In our own analysis (available from the authors upon request), we confirm that these post-election contributions flow largely to the incumbent, suggesting that post-election financial activity could exacerbate imbalances. This is important because, unlike the contribution data, it is impossible to separate the expenditure data into pre- and post-election. Thus, the larger imbalance found on expenditures is likely to be driven, at least in part, by post-election activity.

[7]The p-values reported differ slightly from the ones depicted in Figure 2 of Caughey and Sekhon (2011) because we restrict attention to the subset of districts for which the party of the incumbent is defined, and also because we employ ordinary least squares, whereas they employ a Wilcoxon rank sum test.

[8]As expected, we obtain the same results from a separate analysis where we regress each covariate on lagged incumbency, calculate the residuals, and test for balance on the residuals.

TABLE 1 P-Values from Placebo Tests in Caughey and Sekhon (2011) with and without Controlling for Incumbency

Dependent Variable	Original Specification	Including Dem. Win_{t-1}
Democratic Win $t-1$	.00	–
Democratic % Vote $t-1$	.10	.33
Democratic % Margin $t-1$	.03	.58
Incumbent D1 Nominate	.00	.60
Democratic Incumbent in Race	.00	.58
Republican Incumbent in Race	.00	.44
Democratic # Previous Terms	.08	.74
Republican # Previous Terms	.00	.10
Democratic Experience Adv.	.00	.70
Republican Experience Adv.	.00	.31
Partisan Swing	.00	.24
CQ Rating	.00	.47
Democratic Spending %	.01	.22
Democratic Donation %	.07	.53

Note: These placebo tests cover all those with a reported imbalance in Caughey and Sekhon (2011). Cell entries are p-values for the variable *Democratic* Win_t from linear regressions on the set of races in a 0.5-point window, with robust standard errors. In the column labeled "Original Specification," the only regressor is *Democratic* Win_t. In the column labeled "Including *Dem.* Win_{t-1}," the two regressors are *Democratic* Win_t and *Democratic* Win_{t-1}. For full variable definitions, see Caughey and Sekhon (2011).

between incumbency and each of these covariates. This suggests that focusing on incumbency may be sufficient for detecting similar patterns in other electoral settings: imbalance on incumbency produces imbalance on these other variables as well, and the purported imbalances on these other variables go away once we account for incumbency.[9]

The statistical reason for focusing on incumbency is a concern about multiple testing. If we test for differences between winners and losers in a large enough set of variables, we will eventually find it by chance even if the assumptions underlying RD are in fact met. Future studies may seek to test other variables while applying corrections for multiple testing, but here we focus on the single variable that is purported to be the most problematic and conduct the same battery of tests across many different electoral settings.

The theoretical reason for focusing on incumbency is that it confers electoral benefits in a variety of electoral settings around the world (Ariga 2010; Hainmueller and Kern 2008; Horiuchi and Leigh 2009; Katz and King 1999; Kendall and Rekkas 2012).[10] Of course, in particular settings, other factors may confer systematic electoral advantages: In some local elections, for example, candidates may benefit from belonging to the party controlling a higher-level office; in other settings, being part of a political dynasty may be particularly politically advantageous (e.g., Dal Bó, Dal Bó, and Snyder 2009; Querubin 2011). Unlike these factors, incumbency status is well defined and easily measured in all single-seat electoral systems and is thus a natural attribute to focus on as we look for systematic differences between winners and losers of close elections.

Do Incumbents Disproportionately Win Close Elections?

We analyze data for every partisan, single-winner, plurality/majoritarian electoral setting where data could be collected and assembled. This sample includes national legislative elections in every country that has held competitive plurality elections continuously since at least 1960 and local elections in several politically significant settings. In total, we analyze 20 electoral settings in 10 different countries. The data sets are listed in Table 2; in Appendix A in the Supporting Information (SI), we provide the source of each data set and details on how we handled issues such as redistricting and multiparty competition.[11] We follow Caughey and Sekhon (2011) in choosing a reference party for each setting (e.g., the Democrats in U.S. data sets; the Conservatives in U.K. data sets) and calculating vote margins and incumbency status with respect to that party of interest. The vote margin for the reference party is the difference in vote share between the party of interest and the highest finisher among the other parties. Table 2 reports the number of races in each data set (as well as in the pooled data set) where the margin of victory was less than 10, 2, and 1 percentage points. For example, a bandwidth of 1 includes all elections where the reference party won or lost by a margin of 1 point or less. In a case with only two parties, this would include all cases where the reference party won between 49.5 and 50.5% of the vote.

[9]Put another way, even though we observe imbalances on many covariates, they all tap into a single underlying factor (incumbency) and so are not independent pieces of information.

[10]Though see also Linden (2004); Uppal (2009); Aidt, Golden, and Tiwari (2011); and Klašnja and Titiunik (2013) for evidence of incumbency disadvantage in India and Brazil.

[11]In all settings, we omit cases where the difference in vote share between the first- and third-place party is less than 5 percentage points; this is to avoid complexities emerging from close races involving more than two parties.

TABLE 2 Data and Sample Sizes Analyzed

Setting	Bandwidth 10	2	1	Reference Party
U.S., House of Reps, 1880–2010	5087	1084	567	Democratic
U.S., House of Reps, 1880–1944	3232	731	380	Democratic
U.S., House of Reps, 1946–2010	1855	353	187	Democratic
U.S., Statewide, 1946–2010	2202	498	250	Democratic
U.S., State Legislature, 1990–2010	5953	1204	582	Democratic
U.S., Mayors, 1947–2007	457	108	51	Democratic
Canada, Commons, 1867–2011	2553	576	278	Liberal
Canada, Commons, 1867–1911	759	205	102	Liberal
Canada, Commons, 1921–2011	1794	371	176	Liberal
U.K., Commons, 1918–2010	3414	675	336	Conservative
U.K., Local Councils, 1946–2010	10881	2123	1047	Conservative
Germany, Bundestag, 1953–2009	1260	262	131	CDU/CSU
Bavaria, Mayors, 1948–2009	928	195	87	CSU
France, National Assembly, 1958–2007	872	215	104	Socialist
France, Municipalities, 2008	458	104	59	Left
Australia, House of Reps, 1987–2007	349	73	39	Labor
New Zealand, Parliament, 1949–1987	330	57	27	National
India, Lower House, 1977–2004	1093	222	106	Congress
Brazil, Mayors, 2000–2008	1270	265	143	PMDB
Mexico, Mayors, 1970–2009	4016	801	404	PRI
All Races Pooled	41124	8463	4212	—

Note: See Appendix A in the supporting information for details on each data set. The bandwidths are defined such that a bandwidth of 1 includes all elections where the reference party won or lost by a margin of 1 point or less.

Table 3 assesses whether incumbent parties disproportionately win close elections in a variety of settings. Our basic strategy is to test for an "effect" of winning an election at time t on incumbency status at time $t-1$. We carry out this placebo analysis using three common RD approaches. The "difference-in-means" analysis compares the mean values of the placebo outcome (an indicator for whether the reference party won the previous election) in narrow windows above and below the electoral threshold.[12] "Local linear" analysis similarly tests for a jump in incumbency status at the threshold where the party of interest's vote margin changes from negative to positive, but it does so by fitting linear regressions on each side of the electoral threshold to account for a potential slope of the regression function in the window around the threshold. "Polynomial" does the same thing but with a third-order polynomial regression. For each type of analysis, we summarize the results by reporting the p-value on the test for a jump at the threshold, using italics to signal that the placebo treatment effect is negative, (i.e. that incumbents appear to do *worse*). In SI Appendix B, we present these results graphically and for more specifications. Specifically, in Figures B2–B5, we present the results from the local linear specification for all possible bandwidths between 0.5 and 5. These graphs

[12]The analysis with a bandwidth of 0.5 is equivalent to a test for a difference in the binned means on either side of the threshold in Figure 1. In the RD literature, this is sometimes called a "naive" specification. Despite the benefit of simplicity and transparency, it could produce biased estimates because the potential outcomes are likely correlated with the running variable, even in a small window. For this reason, this specification is only recommended for very small bandwidths where the bias is likely to be negligible. In this particular setting, this bias is likely to lead us to overestimate the success of the incumbent-party in close elections because party performance is positively correlated over time. See Imbens and Lemieux (2008, 624) for a formal discussion of the bias of the difference-in-means estimator in the RD context. They advocate against the difference-in-means estimator in the RD context because it is likely that the bias is "relatively high." Figure B1 in the SI appendix B shows an example of this where in our pooled sample of all close races, the difference-in-means estimator is biased even within a bandwidth of 1 percentage point because it ignores the positive slope within the bin.

TABLE 3 Placebo Tests: p-values for "Effect" of Party Winning at Time *t* on Party Winning at Time *t* − 1

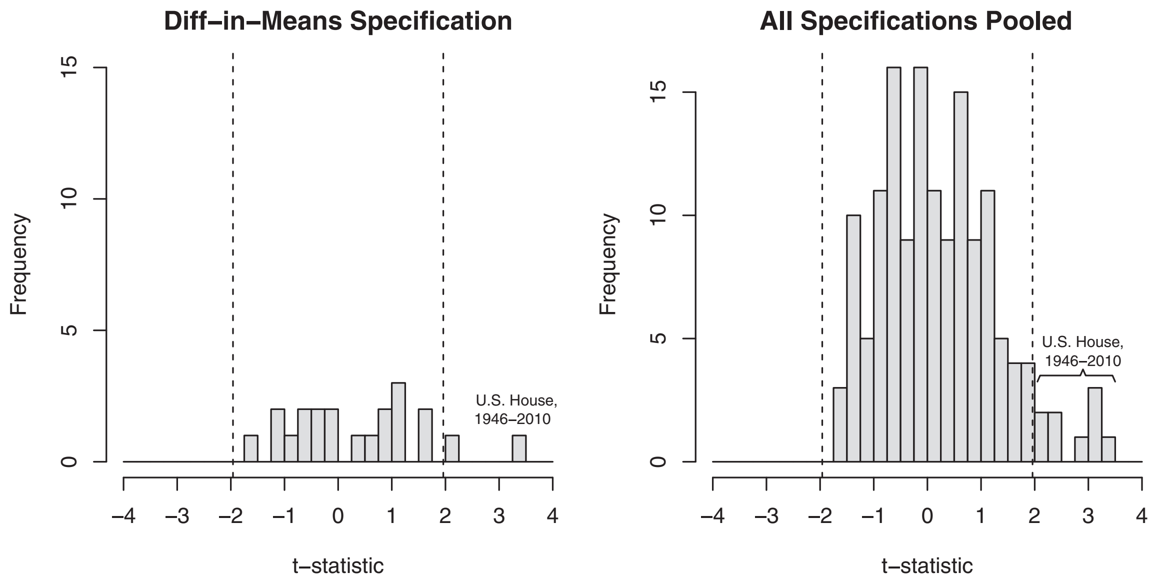
Setting	**Diff-in-Means**		**Local Linear**			**Polynomial**	
***Bandwidth* =**	***0.5***	***1***	***1***	***2***	***5***	***5***	***10***
U.S., House of Reps, 1880–2010	0.11	0.07	0.46	0.30	0.33	0.30	0.33
U.S., House of Reps, 1880–1944	*0.70*	*1.00*	*0.59*	*0.36*	*0.90*	*0.48*	*0.62*
U.S., House of Reps, 1946–2010	0.00	0.00	0.04	0.00	0.07	0.00	0.02
U.S., Statewide, 1946–2010	*0.55*	*0.79*	*0.43*	*0.38*	*0.56*	*0.50*	*0.10*
U.S., State Legislature, 1990–2010	*0.37*	0.52	*0.32*	0.95	0.59	*0.78*	*0.77*
U.S., Mayors, 1947–2007	—	0.96	—	0.81	0.88	0.37	*0.62*
Canada, Commons, 1867–2011	*0.29*	*0.50*	*0.32*	*0.18*	*0.09*	*0.59*	*0.17*
Canada, Commons, 1867–1911	*0.59*	*0.22*	0.81	*0.21*	*0.19*	*0.60*	*0.18*
Canada, Commons, 1921–2011	*0.30*	*0.88*	*0.18*	*0.39*	*0.17*	*0.71*	*0.35*
U.K., Commons, 1918–2010	0.33	0.09	0.59	0.61	0.08	*0.92*	0.12
U.K., Local Councils, 1946–2010	0.24	0.06	0.44	0.27	0.22	0.17	0.68
Germany, Bundestag, 1953–2009	0.71	0.54	0.79	0.48	*1.00*	0.74	0.84
Bavaria, Mayors, 1948–2009	*0.13*	*0.38*	*0.21*	*0.39*	*0.16*	*0.19*	*0.30*
France, National Assembly, 1958–2007	0.27	0.79	0.33	0.55	0.53	0.47	0.23
France, Municipalities, 2008	—	0.31	—	0.37	0.14	0.52	0.24
Australia, House of Reps, 1987–2007	—	—	—	1.00	*0.55*	0.50	*0.92*
New Zealand, Parliament, 1949–1987	—	—	—	—	0.75	*0.86*	0.69
India, Lower House, 1977–2004	0.49	0.38	*0.54*	*0.98*	0.20	*0.97*	0.86
Brazil, Mayors, 2000–2008	*0.81*	*0.81*	*0.61*	*0.58*	0.78	*0.64*	0.97
Mexico, Mayors, 1970–2009	*0.69*	*0.96*	*0.39*	*0.68*	*0.93*	*0.93*	*0.60*
All Races Pooled	0.22	0.02	0.92	0.59	0.16	0.46	0.75

Note: Each entry gives the p-value of a two-tailed test of the hypothesis that the coefficient on *Treatment* is zero. Results not shown if there are insufficient data points within a given bandwidth, to avoid biased or uninformative inferences. Sample size cutoffs are 40, 60, and 100 for difference-in-means, local linear, and polynomial. Results in italics indicate that the point estimate is the opposite of what one would expect if incumbents disproportionately win close elections. Robust standard errors are used in all cases. Standard errors clustered by state-year for U.S. statewide offices.

also present the point estimates for readers interested in interpreting the substantive size of the point estimates directly and show that the results are robust across many specifications.

As expected, our tests uncover the imbalance in the U.S. House in the post-World War II period (row 3). Previous papers have focused on the difference-in-means specification, and we replicate this result for other RD specifications as well. However, for the U.S. House in the previous period as well as for the U.S. House in the entire period since 1880, we fail to find evidence of incumbent advantages in any specification at the .05 level. Turning to the other U.S. contexts (i.e., statewide offices since 1946, state legislatures since 1990, and mayors since 1947), we find no evidence of an advantage for the incumbent party in any specification. This finding is particularly interesting given that existing explanations for incumbents' disproportionate success in the postwar U.S. House would seem to apply at least as strongly to these other contexts. Outside the United States, we similarly fail to find any evidence of an advantage to incumbent party candidates. Out of 96 tests shown for non-U.S. data, we do not find a single p-value below .05. When we pool all of the data into a single data set (bottom row of the table), we similarly find no evidence of incumbent advantages. The one case where the p-value is below .05 is the difference-in-means analysis with a bandwidth of 1, but a closer investigation of this reveals that the difference-in-means estimate is highly biased upward, since it ignores the strong positive slope within the bandwidth (see Figure B1 in the SI appendix, which plots the relationship between lagged incumbency and the margin of victory for these close races and shows that even within a 1 percentage point bandwidth, the difference-in-means estimator provides a poor approximation to the limits from below and above of the regression functions toward the threshold). Given

FIGURE 2 T-values for "Effect" of Party Winning at Time t on Party Winning at Time t − 1



this bias, we do not view this estimate as evidence of imbalance.[13]

Figure 2 provides a graphical summary of the results in Table 3. In the left panel, we plot the histogram of the t-statistics of the tests in the first column of Table 3—difference-in-means estimates of the difference in lagged victory rate between close winners and losers for a bandwidth of 0.5. The t-statistics are evenly distributed around 0 except for a single outlier above 3: the U.S. House in the post-World War II period. In the right panel, we include all of the (nonpooled) tests from Table 3. Again, the distribution appears to be roughly unimodal about 0, except for a right tail; *every one* of the t-statistics greater than 1.96 comes from the U.S. House in the post-World War II period. We present these results graphically and for many more specifications in SI Appendix B (Figures B2 and B4).

As noted above, our placebo tests focus on (lagged) incumbency because our analysis in Table 1 suggests that incumbency accounts for most of the imbalances reported in existing studies for the U.S. House. It is good practice, however, to check for balance in the lagged running variable (Imbens and Lemieux 2008), that is, the vote margin in the previous race. Table 4 reports results of the same tests using the same format as Table 3, where the outcome is the lagged vote margin rather than lagged incumbency status. The difference-in-means analysis shows imbalance in the U.S. House only at the 1-point bandwidth for the post–World War II period; in no setting is there consistent evidence of imbalance. Again, we present these results graphically and for many more specifications in SI Appendix B (Figures B3 and B5). Histograms of test statistics are displayed in Figure 3 and indicate a pattern similar to the one in Figure 2: t-statistics appear to be drawn from a unimodal density centered about 0.

In Table 5, we report the results of additional analyses based on the density test suggested by McCrary (2008). In these tests, we assess whether the density of incumbent party candidate vote share is smooth near the electoral threshold. We first separate each data set according to whether the party of interest previously won the seat ("Incumbent" versus "Nonincumbent") and carry out the McCrary test separately on each series, restricting attention to cases where the margin of victory was within 10 percentage points. If incumbents disproportionately win close elections, we would expect a break in the density of the vote margin at 0—a jump up for the sample of elections in which the party of interest held the seat and a drop down for the sample of elections in which the party of interest did not hold the seat. We do not generally find this pattern; even the results for the U.S. House in the post–World War II period are only borderline significant for the "Incumbent" series. We then recombine the two subsets while flipping the sign of the vote margin for the cases in which the party of interest was not the incumbent; for this combined data set, we would expect a bulge in the density where the adjusted margin is slightly above 0,

[13]In fact, if party performance is correlated over time, a difference-in-means test should yield a significant result at any bandwidth given sufficient data, even if incumbents have no special advantages in close elections.

TABLE 4 Placebo Tests: p-values for "Effect" of Party Winning at Time *t* on Party Vote Margin at Time *t* − 1

Setting	Diff-in-Means		Local Linear			Polynomial	
Bandwidth =	**0.5**	**1**	**1**	**2**	**5**	**5**	**10**
U.S., House of Reps, 1880–2010	0.21	0.15	0.81	0.51	0.37	0.77	0.81
U.S., House of Reps, 1880–1944	0.91	*0.85*	*0.77*	*0.46*	*0.95*	*0.39*	*0.58*
U.S., House of Reps, 1946–2010	0.15	0.04	0.63	0.16	0.21	0.29	0.41
U.S., Statewide, 1946–2010	*0.84*	0.69	*0.81*	*0.82*	*0.98*	0.97	*0.29*
U.S., State Legislature, 1990–2010	0.75	0.78	0.92	*0.91*	*0.91*	*0.89*	*0.59*
U.S., Mayors, 1947–2007	—	*0.11*	—	*0.22*	*0.42*	*0.09*	*0.10*
Canada, Commons, 1867–2011	*0.12*	*0.31*	*0.13*	*0.10*	*0.06*	*0.29*	*0.08*
Canada, Commons, 1867–1911	*0.26*	*0.17*	*0.38*	*0.27*	*0.08*	*0.53*	*0.12*
Canada, Commons, 1921–2011	*0.21*	*0.51*	*0.20*	*0.17*	*0.17*	*0.35*	*0.19*
U.K., Commons, 1918–2010	0.16	0.11	0.65	0.43	0.58	0.67	0.46
U.K., Local Councils, 1946–2010	0.10	0.02	0.33	0.12	0.40	0.08	0.35
Germany, Bundestag, 1953–2009	0.95	*0.45*	0.50	*0.81*	*0.29*	0.98	*0.37*
Bavaria, Mayors, 1948–2009	*0.10*	*0.39*	*0.12*	*0.30*	*0.10*	*0.23*	*0.26*
France, National Assembly, 1958–2007	*0.57*	*0.39*	*0.54*	*0.26*	*0.76*	*0.34*	*0.92*
France, Municipalities, 2008	—	0.46	—	0.83	0.11	0.92	0.48
Australia, House of Reps, 1987–2007	—	—	—	0.49	0.30	0.36	0.18
New Zealand, Parliament, 1949–1987	—	—	—	—	0.09	*0.77*	0.31
India, Lower House, 1977–2004	*0.77*	0.78	*0.40*	*0.78*	0.21	*0.88*	0.89
Brazil, Mayors, 2000–2008	*0.47*	*0.77*	*0.25*	*0.33*	0.52	*0.32*	0.95
Mexico, Mayors, 1970–2009	*0.99*	*0.77*	0.83	0.98	*0.35*	0.73	*0.42*
All Races Pooled	0.46	0.25	*0.95*	*0.88*	0.95	*0.95*	*0.50*

Note: See text for explanation of tests and notes to Table 3 for details on presentation.

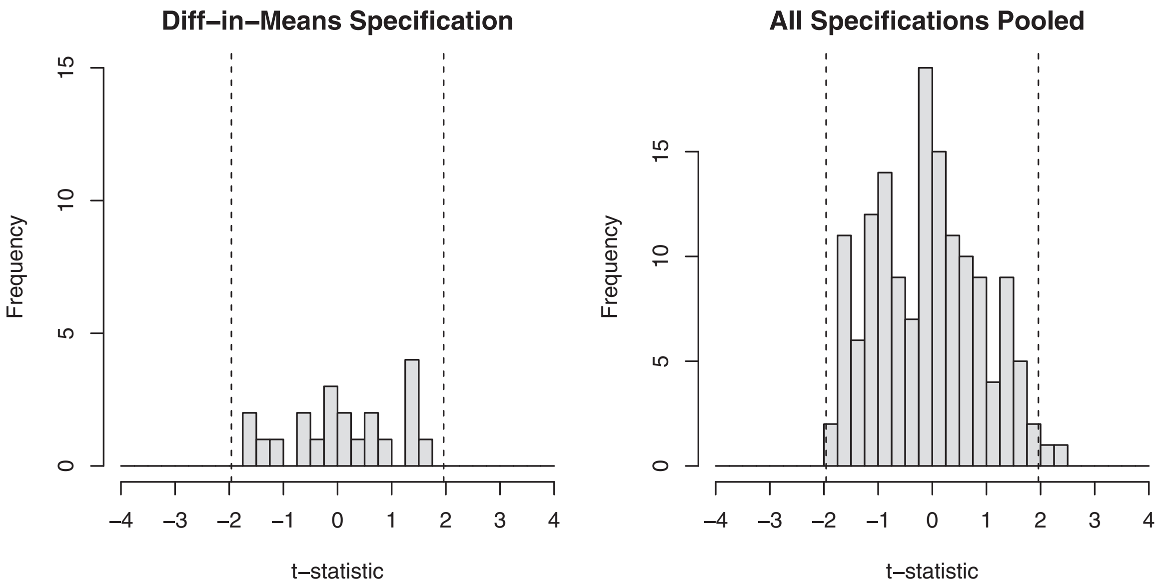
indicating that the party of interest is likely to narrowly lose when it previously lost and likely to narrowly win when it previously won. As indicated by Table 5, we cannot reject the null of no density jump for any setting except the U.S. House after 1946.

What Mechanisms Could Lead to Imbalance in Electoral RD Designs?

The analysis in the previous section indicates that the apparent dominance of incumbent party candidates is limited to the U.S. House in the post–World War II period. What does this mean for the use of electoral RD designs? The most optimistic conclusion is that the disproportionate rate of success among incumbents in close House elections is the result of statistical chance, which would indicate no fundamental problem for electoral RD analysis (although researchers applying an RD to the U.S. House need to take special care). Other interpretations are possible, however. For example, one could conclude that some class of candidates is able to precisely control electoral outcomes in many settings, but that this advantaged class varies across settings. If so, we might find imbalance in incumbency status only in the U.S. House (and only in the post-WWII period), even though the assumptions behind the electoral RD design are violated more widely.

In order to clarify the significance of the imbalances in the postwar U.S. House, we briefly discuss the theoretical mechanisms through which incumbents (or other structurally advantaged candidates) could exert fine control over the outcomes of close elections. Along the way, we assess the plausibility of those mechanisms in the case of the U.S. House. In the end, we conclude that none of the current explanations for the imbalance observed in the U.S. House are satisfying. This suggests that this imbalance might be the result of chance. Nonetheless, researchers must think carefully about these potential mechanisms, whether they are present in a particular electoral setting, and whether they might bias estimates arising from future RD designs. We also use this discussion to motivate our next section, which provides a set of best practices—both theoretical and empirical—that future researchers should

FIGURE 3 T-values for "Effect" of Party Winning at Time t on Party Vote Margin at Time t − 1



employ when implementing RD designs in electoral settings.

Explanations for systematic advantages of incumbents (or other advantaged candidates) in close elections can be crudely divided into two categories: those that focus on *pre-election* behavior, like the campaign efforts that Caughey and Sekhon (2011) discuss, and those that focus on *post-election* behavior, including the processing of ballots and the recount process. We consider each type of explanation in turn.

There are several theoretical requirements for any *pre-election* explanation for imbalance. For example, advantaged candidates must have access to additional (but costly) resources that they only employ when necessary, they would have to obtain extremely precise information about their expected vote share, and the opposing campaign must lack the ability or willingness to do these same things. Here, we focus on the most salient of these requirements: information.

Recall that the imbalance observed in the U.S. House is present for only a tiny window around the electoral threshold, where the Democratic win margin was less than 0.5 percentage points (i.e., those elections where the Democratic two-party vote percentage is between 49.75 and 50.25). If strategic campaigning or other pre-election behaviors explain this imbalance, then incumbent behavior must vary significantly across small changes in the expected election result. Specifically, their behavior would have to be systematically different in scenarios where they would expect vote percentages between 49.75 and 50, compared to other scenarios where they would expect vote percentages in the bins immediately outside of this range. For example, incumbents would behave differently if they expect to receive 49.9% of the two-party vote as opposed to 49.7 or 50.1%. Perhaps at 49.9, incumbents exert extra effort in an attempt to win, but at 49.7, they know the cause is lost so they do not bother, and at 50.1, they rest assured of victory and similarly do not bother exerting extra effort. Of course, this explanation assumes that incumbents can reasonably distinguish between situations where they expect to receive 49.7, 49.9, and 50.1% of the vote. In SI Appendix C, we provide a theoretical model of campaign effort and show that incumbent campaigns would have to predict their vote shares within approximately one-quarter of 1 percentage point (at most), on average, in order for pre-election behavior to explain the pattern of imbalance that we observe in the U.S. House.

The realities of political polling and congressional campaigns cast serious doubt on the ability of candidates to obtain such precise expectations. Enos and Hersh (2013) provide evidence on the precision of campaign expectations by surveying Democratic candidates and campaign operatives in the run-up to the 2012 general election. On average, campaign workers mispredict their vote share by 8 percentage points, and this lack of precision does not vary meaningfully across the status of the campaign worker (candidates and high-level managers are no better than volunteers and lower-level workers), the competitiveness of the race, the time until the election, or incumbent versus challenger campaigns.

TABLE 5 McCrary (2008) Tests: p-values for Null Hypothesis of Equal Density on Opposite Sides of the Threshold

Setting	Incumbent	Non-incumbent	Pooled
U.S., House of Reps, 1880–2010	0.80	0.85	0.95
U.S., House of Reps, 1880–1944	*0.60*	*0.57*	*0.38*
U.S., House of Reps, 1946–2010	0.07	0.18	0.05
U.S., Statewide, 1946–2010	*0.43*	*0.47*	*0.26*
U.S., State Legislature, 1990–2010	*0.83*	*0.42*	*0.41*
U.S., Mayors, 1947–2007	*0.76*	0.13	0.39
Canada, Commons, 1867–2011	*0.34*	*0.62*	*0.23*
Canada, Commons, 1867–1911	0.65	*0.14*	*0.38*
Canada, Commons, 1921–2011	*0.25*	0.59	*0.76*
U.K., Commons, 1918–2010	0.44	0.07	0.10
U.K., Local Councils, 1946–2010	*0.73*	0.32	0.46
Germany, Bundestag, 1953–2009	*0.49*	0.33	0.64
Bavaria, Mayors, 1948–2009	*0.26*	0.83	0.93
France, Natl Assembly, 1958–2007	*0.62*	0.03	0.12
France, Municipalities, 2008	—	0.91	0.10
Australia, House of Reps, 1987–2007	*0.72*	*0.13*	*0.13*
New Zealand, Parliament, 1949–1987	0.40	1.00	0.78
India, Lower House, 1977–2004	*0.79*	0.40	0.58
Brazil, Mayors, 2000–2008	0.45	*0.37*	0.83
Mexico, Mayors, 1970–2009	*0.94*	*0.63*	*0.85*
All Races Pooled	*0.81*	0.42	0.62

Note: See text for explanation of test and notes to Table 3 for details on presentation.

For the five "toss-up" U.S. House races where Enos and Hersh (2013) surveyed the incumbent campaign, the operatives mis-predicted the election result by 10 percentage points, on average. Statistical models reveal similar levels of uncertainty about the outcomes of close elections. Klarner (2008) generates race-by-race predictions for the two-party vote share in every contested House election in 2008. On average, for contested races, these predictions miss the actual election result by 4.3 percentage points, and the average error exceeds 6 percentage points for the most competitive races. Likewise, the final poll or even the average of many late polls in a close U.S. House race in 2012, on average, missed the actual election result by about 2 percentage points.[14] With this information available, then, congressional candidates can hardly tell the difference between situations where they are likely to lose narrowly and those where they are likely to win narrowly. In fact, because election outcomes are so uncertain, modern campaign managers and consultants often aim for *52%* of the two-party vote.[15] We do not know how they decided upon this magic number, but the fact that these campaigns do not target the actual threshold suggests that campaign activity is unlikely to explain the precise imbalance.

Post-election explanations for imbalance—revolving around court cases, recounts, post-election fraud, and so on—are theoretically more plausible. In these cases, candidates might know exactly when to exert costly effort because the initial vote count is public. Whether or not incumbent candidates (or some other class of candidates) can disproportionately win these battles then depends on the specifics of the particular setting. In the case of the U.S. House, Caughey and Sekhon (2011) rule out these explanations after finding that while recounts occur frequently in close races, they rarely reverse the initial result. This is consistent with the idea that incumbent party candidates and challengers both bring substantial resources to election contests and thus incumbents cannot dominate at the

[14]We conducted this analysis ourselves by collecting all of the polls available through Real Clear Politics.

[15]This was relayed to us in private correspondence with a campaign consultant.

recount stage.[16] Other post-election mechanisms would include more flagrantly illegal behavior, such as altering precinct-level vote tallies after all of the results have been counted. For such a mechanism to account for incumbent dominance in very close U.S. House races, electoral manipulation would have to be widespread, and this type of outright fraud is thought to be rare in this setting and time period (Lehoucq 2003). Moreover, we lack an explanation for why such behavior would be present in postwar House elections but absent in the prewar House and in postwar elections for state legislatures and statewide offices.

In sum, we find existing post-election and pre-election explanations of observed imbalances in close U.S. House races to be fairly implausible. Outside of structural advantages to incumbents (or some other class of candidates) in manipulating electoral tallies after the election or in winning legal challenges, there exists no convincing theoretical reason to expect close winners and losers of a large election to differ systematically. The implausibility of the mechanisms that have been suggested to explain imbalance in the postwar U.S. House suggests that the success of incumbent party candidates in very close elections likely reflects statistical chance. To be sure, if we look at close elections in the postwar U.S. House in isolation, we observe a degree of incumbent party success that appears unlikely to have arisen randomly.[17] However, given a large number of electoral settings, it is likely that this degree of imbalance would emerge in one of them simply by chance. The analysis in this article suggests that the postwar U.S. House may be that exceptional setting in which imbalance arose by chance.[18] Of course, this does not preclude the possibility that future work might uncover a more compelling explanation for imbalance in the U.S. House that could lead us to revise this conclusion.

Recommendations for Future Researchers

In examining the observed imbalance in the U.S. House, as well as in presenting our tests for other electoral offices, we have touched upon the techniques that we believe researchers should employ when validating the RD design in applied settings. The fact that we fail to find problems in numerous electoral settings does not excuse researchers from defending the identification assumptions of their empirical strategies with both theory and data. The burden of proof is on the researcher to justify her assumptions and subject them to rigorous testing. A key advantage of the RD design is that it lends itself to numerous, transparent tests that follow directly from the identification assumptions. In this section, we propose a set of best practices for future researchers. We do not focus on the technical details of the RD design, which have already been clearly laid out in, for example, Hahn, Todd, and Van der Klaauw (2001), Lee (2008), and Imbens and Lemieux (2008).

To ensure that RD results are both valid and robust, we propose a three-step process. Researchers employing the RD design should engage in the following:

(1) Consider theoretical mechanisms that could produce sorting around the discontinuity.
(2) Evaluate balance on pre treatment covariates and especially on the lagged outcome variable, focusing on the presence or absence of substantively large imbalances in characteristics that might be related to the mechanisms that could produce sorting. These tests should employ the same specifications as those employed to estimate the effects of interest, and these specifications should account for the potential relationship between the running variable and the outcome variable.
(3) Present estimates at a number of alternative bandwidths and specifications.

We now discuss these three steps in detail.

[16] However, all of the four reversals identified and discussed by Caughey and Sekhon (2011) benefited the incumbent party, so recounts may explain some of the observed imbalance. If future work demonstrates that the imbalance in the House is primarily explained by recounts and court cases, there is a workable solution for applied researchers. If the initial vote tally is well behaved but incumbents disproportionately prevail in recounts, then one can employ a "fuzzy" RD design in which the initial vote tally provides an instrument for the final election result. Note that this requires the usual fuzzy RD assumptions, including monotonicity and excludability (see, e.g. Hahn, Todd, and Van der Klaauw 2001). The fuzzy RD also changes the estimand to the local average treatment effect for compliers, but in practice this estimand will be very close to the one from the sharp RD if recounts rarely reverse the initial vote result and therefore the rate of compliance is very high. We should also point out that data on the initial tallies may be difficult to collect in many cases.

[17] We cannot say with precision how unlikely this is. With some specifications, the imbalance appears to be extremely unlikely (e.g., $p < .001$), but for other specifications, the imbalance is only moderately unlikely (e.g., $p = .07$). For obvious reasons, we should not focus only on the specification with the lowest p-value.

[18] For example, across 20 independent settings under the null hypothesis, there is a 64% chance of obtaining at least one p-value below .05 and an 18% chance of obtaining at least one p-value below .01.

Evaluating the RD Assumption Theoretically

While the RD design is an extremely valuable tool for estimating electoral effects, it is not a panacea. The assumptions of the design are often weaker than those of other designs, but they are not guaranteed to hold. For example, if an electorate is small enough that relevant actors could closely predict or manipulate the vote tally, then the RD assumptions would be invalid. For this reason, the RD design should probably not be used to study the effects of judicial or legislative decisions, where strategic voting, endogenous agendas, or bargaining could lead to systematic differences between successful and unsuccessful motions. As a case in point, McCrary (2008) demonstrates that roll-call votes in the U.S. House exhibit sorting around the majority threshold, indicating that such votes do not generate a quasi-random assignment of policy decisions. Similarly, in an electoral setting where all close elections were ultimately decided in courtrooms that often reversed the initial counting of ballots, one could only assume that election winners and losers were comparable if one were willing to assume that the legal process was not systematically biased toward one type of candidate. For these reasons, a researcher must first provide theoretical justification for her design before examining the data. In any new electoral setting, the researcher should ask the following questions: Is the assumption of the RD design, that potential outcomes are smooth at the electoral threshold, defensible a priori? Are there substantive features of this electoral setting that could easily lead the bare winners to be systematically different from the bare losers?

Validating the RD Assumption Empirically

Having considered possible threats to the validity of the RD design theoretically, researchers should then test their assumptions to the extent possible. At a minimum, they should conduct tests for placebo effects of the treatment on the lagged outcome variable when possible. We also highly recommend that researchers show additional placebo tests for the lagged running variable, lagged treatment variable, and other pretreatment covariates, if available. These placebo tests should mimic, as closely as possible, the specifications used to estimate the primary quantities of interest. We discuss the choice of specifications below. Additionally, graphs and/or formal tests for sorting based on McCrary (2008) would further bolster readers' confidence in the underlying assumptions and results (see also Imbens and Lemieux 2008 and Lee and Lemieux 2010 for checklists of tests).

In performing these placebo tests, researchers and consumers should keep in mind the multiple testing problem. Testing for imbalance on many variables makes it likely that some tests will be statistically significant by random chance. Imbalance should therefore be assessed based on the substantive size of the imbalance, and not only on the statistical significance of the balance test. For example, in our analyses, our failure to reject the null was not a product of large standard errors. The substantive levels of imbalance are quite small; see, for example, Figures B2-B5 in the SI appendix.[19] In addition to its value for assessing the presence of imbalances, this is important in considering the sensitivity of analyses performed on the data; the larger the size of the imbalance, the more sensitive estimates are likely to be. Moreover, multiple testing adjustment could be used to adjust the p-values from the placebo checks to control the family-wise error rate.[20]

Demonstrating the Stability of RD Findings

Finally, researchers should assess the extent to which their effect estimates are sensitive to specification. As with many empirical approaches, RD designs leave researchers with some degrees of freedom that can lead to specification searching and false-positive results. To mitigate these concerns, the researcher should show results for many different bandwidths and specifications (e.g., difference-in-means, local linear, polynomial) and also explore sensitivity to the inclusion of pretreatment covariates. The particular specifications should also be justified with theory and data. For example, a difference-in-means approach with a large bandwidth would likely lead to a large bias if the slope of the control function is nonzero, and a high-order polynomial approach with a tiny bandwidth would likely be imprecise and unreliable. Moreover, a local linear specification might be biased if the true regression function is nonlinear within the estimation window.

The researcher should also present her data visually in a transparent way that clarifies the appropriateness of the specification and the sensitivity of the results to changes in the specification. At a minimum, we recommend that researchers show the "main" RD graph that

[19]In our pooled analysis, moreover, we can statistically reject substantively tiny levels of imbalance.

[20]The family-wise error rate is the probability that at least one of the true hypotheses in a family of tested hypotheses is rejected. An attractive methodology for this is the free step-down resampling procedure (Anderson 2008; Westfall and Young 1993). This method is typically more powerful than the Bonferroni correction.

visualizes the relationship between the outcome and the running variable in the benchmark estimation window. Binned local averages should be used to assess the size of the discontinuity and the empirical shape of the regression functions on both sides of the threshold. We also recommend that researchers superimpose predicted values from a flexible control function fitted on both sides of the threshold to help assess the appropriate specification. We also recommend graphs like those we present in the supporting information, in which the point estimate for a given specification (e.g., local linear) is plotted across a large range of plausible bandwidths that are consistent with the specification checks, along with 95% confidence intervals (see also Lee and Lemieux 2010).

RD Estimates with Imbalance

How should researchers proceed if they want to estimate electoral effects in the postwar U.S. House or another setting where imbalance is present? So long as they have ruled out plausible theoretical mechanisms for the imbalance, researchers hoping to estimate electoral effects in the modern U.S. House should proceed in a similar manner to researchers who discover chance imbalances in experimental data.[21] One might still be able to draw inferences from imbalanced experiments given additional assumptions and covariate adjustment.[22] One could adjust for imbalance by including lagged incumbency and other pretreatment variables as covariates in the RD analysis or by preprocessing the data through matching or reweighting before conducting the RD analysis. Alternatively, researchers might consider a "donut" RD design (Almond and Doyle 2011; Barreca et al. 2011), where they exclude the small sample of very close elections where imbalance exists.[23] It is important to emphasize that all of these fixes require additional assumptions that need to be justified, and extraordinary care is required in order to generate inferences given the presence of imbalances. Even if there is something fundamentally problematic about the RD assumptions in the U.S. House, the RD design may still be the best of all imperfect methods for estimating electoral effects in this important setting, and careful RD analysis may still produce better estimates than we could have otherwise obtained with other empirical strategies. As Caughey and Sekhon write, even in the case of estimating electoral effects in recent U.S. House elections, the RD design appears to be the best option: "Nevertheless, a comparison of the Lee RD estimator with traditional regression approaches to the incumbency advantage reveals that RD relies on weaker assumptions" (2011, 405).

Conclusion

Our results should not induce complacency about the validity of RD designs in close elections. However, they should place the documented imbalances in U.S. House elections in the proper context. Our perception is that papers showing disproportionate incumbent successes in the U.S. House (particularly Caughey and Sekhon 2011) have been highly influential among political scientists interested in estimating electoral effects. Absent careful analysis of other electoral contexts, one might conclude that there is something fundamentally problematic about the use of RD to study electoral effects. Evidence of imbalance in the U.S. House may have even made some scholars suspicious of all RD-based studies, to the point where they lend more credence to other approaches. The RD imbalance literature, to our reading, never intended this reaction. Indeed, Caughey and Sekhon point out that the RD design "still makes weaker assumptions than the usual model-based alternatives" (2011, 406). We agree strongly with this sentiment, and we hope that the validity tests presented in this article make it clear that the RD design is broadly applicable.

To our knowledge, this article provides the most thorough and extensive assessment to date of the validity of the regression discontinuity design in electoral settings. Across more than 40,000 closely contested races in many different electoral settings, we find no systematic evidence of sorting or imbalance around electoral thresholds. Conditional on being in a very close election, incumbents are no better at winning than challengers. We hope that these results will bolster confidence in estimates of electoral effects that arise from RD designs, so long as researchers exercise the appropriate level of rigor. We combine this

[21] Of course, we cannot be as certain that this imablance has arisen by chance as we could in an experimental setting, and this difference warrants additional prudence.

[22] See Rubin (1973, 1979, 2009), Schochet (2010), and Miratrix, Sekhon, and Yu (2013) for discussions of when and how valid inferences can be drawn from imbalanced experimental data.

[23] To be clear, the "donut" has been developed specifically for cases in which there are strong a priori reasons to expect heaping in the running variable. However, one might imagine cases in which the threshold for recounts is known to the researcher, and the researcher believes that recounts are not random. In such a case, the researcher might employ a "donut" approach. Nonetheless, it is important to keep in mind that excluding data near the threshold is usually undesirable, since these are the most useful data points in a typical RD. Nevertheless, if these data points are suspect, robustness to their exclusion is a good sign. With sufficient data farther away, but still close, to the threshold, one might still extrapolate to the discontinuity with these points removed.

analysis with a consideration of theoretical mechanisms through which the RD assumptions may be violated, arguing that in the case of the U.S. House, a plausible mechanism has not yet been proposed; this further strengthens our confidence in the validity of using RD to estimate electoral effects.

To aid in this rigor, we have used our analyses as an opportunity to present our recommendations on best practices for applied RD users. When considering the use of the RD design in applied work, researchers should begin by considering theoretical reasons for the violation of the RD assumption. If the assumption appears theoretically plausible, researchers should perform a battery of balance tests on pretreatment covariates and lagged values of the outcome variable, using the same specifications as the analysis on the outcome variable. In performing these tests, researchers should keep in mind that large numbers of tests will lead to some false positives, and so they should place special emphasis on the substantive size of any observed imbalances and/or adjust for multiple testing explicitly. Finally, we recommend that researchers present graphical evidence to support the appropriateness of the specifications used to estimate the effects on the outcome variable of interest and report the estimated effects across a large number of bandwidths and specifications of the running variable.

The RD design provides the opportunity for researchers to assess electoral effects under unusually weak assumptions that mitigate issues of model dependency and omitted variables in all but the most unusual cases. The best practices we propose in this article should allow researchers to apply the RD design, when justified through theory and validation, with the confidence that they have addressed possible problems of imbalance in their data. Though the RD assumption may not always hold, it continues to offer the most plausible, least model-dependent estimates for a variety of electoral effects across numerous electoral settings.

References

Ade, Florian, and Ronny Freier. 2011. "Divided Government versus Incumbency Externality Effect: Quasi-Experimental Evidence on Multiple Voting Decisions." DIW Berlin Discussion Paper No. 1121.

Aidt, Toke, Miriam A. Golden, and Devesh Tiwari. 2011. "Incumbents and Criminals in the Indian National Legislature." Unpublished manuscript, University of Cambridge.

Almond, Douglas, and Joseph J. Doyle. 2011. "After Midnight: A Regression Discontinuity Design in Length of Postpartum Hospital Stays." *American Economic Journal: Economic Policy* 3(3): 1–34.

Anderson, Michael L. 2008. "Multiple Inference and Gender differences in the effects of Early Intervention: A Rreevaluation of the Abecedarian, Perry Preschool, and Early Training Projects."*Journal of the American Statistical Association* 103(484): 1481–95.

Ariga, Kenichi. 2010. "Entrenched Incumbents, Irresponsible Parties? Comparative Analysis of Incumbency Advantage across Different Electoral Systems." Ph.D. dissertation, University of Michigan.

Barreca, Alan I., Melanie Guldi, Jason M. Lindo, and Glen R. Waddell. 2011. "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification." *Quarterly Journal of Economics* 126(4): 2117–23.

Boas, Taylor C., and F. Daniel Hidalgo. 2011. "Controlling the Airwaves: Incumbency Advantage and Community Radio in Brazil." *American Journal of Political Science* 55(4): 869–85.

Broockman, David E. 2009. "Do Congressional Candidates Have Reverse Coattails? Evidence from a Regression Discontinuity Design." *Political Analysis* 17(4): 418–34.

Butler, Daniel Mark. 2009. "A Regression Discontinuity Design Analysis of the Incumbency Advantage and Tenure in the U.S. House." *Electoral Studies* 28(1): 123–28.

Caughey, Devin, and Jasjeet S. Sekhon. 2011. "Elections and the Regression Discontinuity Design: Lessons from Close U.S. House Races, 1942–2008." *Political Analysis* 19(4): 385–408.

Cellini, Stephanie Riegg, Fernando Ferreira, and Jesse Rothstein. 2010. "The Value of School Facility Investments: Evidence from a Dynamic Regression Discontinuity Design." *Quarterly Journal of Economics* 125(1): 215–61.

Dal Bó, Ernesto, Pedro Dal Bó, and Jason Snyder. 2009. "Political Dynasties." *Review of Economic Studies* 76(1): 115–42.

Dell, Melissa. 2012. "Trafficking Networks and the Mexican Drug War."Unpublished manuscript, Harvard University.

DiNardo, John, and David S. Lee. 2004. "Economic Impacts of New Unionization on Private Sector Employers: 1984–2001." *Quarterly Journal of Economics* 119(4): 1383–1441.

Eggers, Andrew C., and Jens Hainmueller. 2009. "MPs for Sale? Returns to Office in Postwar British Politics." *American Political Science Review* 103(4): 513–33.

Enos, Ryan, and Eitan Hersh. 2013. "Elite Perceptions of Electoral Closeness: Fear in the Face of Uncertainty or Overconfidence of True Believers." MPSA Annual Meeting Paper.

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

Folke, Olle, and James M. Snyder Jr. 2012. "Gubernatorial Midterm Slumps." *American Journal of Political Science* 56(4): 931–48.

Gagliarducci, Stefano, and M. Daniele Paserman. 2012. "Gender Interactions within Hierarchies: Evidence from the Political Arena." *Review of Economic Studies* 79(3):1021–52.

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): 326–39.

Grimmer, Justin, Eitan Hirsh, Brian Feinstein, and Daniel Carpenter. 2012. "Are Close Elections Random?" Working paper.

Hahn, Jinyong, Petra Todd, and Wilbert Van derKlaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica* 69(1): 201–09.

Hainmueller, Jens, and Holger Lutz Kern. 2008. "Incumbency as a Source of Spillover Effects in Mixed Electoral Systems: Evidence from a Regression-Discontinuity Design." *Electoral Studies* 27(2): 213–27.

Horiuchi, Yusaku, and Andrew Leigh. 2009. "Estimating Incumbency Advantage: Evidence from Multiple Natural Experiments." Unpublished manuscript, Dartmouth College.

Imbens, Guido W., and Thomas Lemieux. 2008. "Regression Discontinuity Designs: A Guide to Practice." *Journal of Econometrics* 142(2): 615–35.

Katz, Jonathan N., and Gary King. 1999. "A Statistical Model for Multiparty Electoral Data." *American Political Science Review* **(*): 15–32.

Kendall, Chad, and Marie Rekkas. 2012. "Incumbency Advantages in the Canadian Parliament." *Canadian Journal of Economics/Revue Canadienne d'Économique* 45(4): 1560–85.

Klarner, Carl. 2008. "Forecasting the 2008 U.S. House, Senate and Presidential Elections at the District and State Level." *PS: Political Science & Politics* 41(4): 723–28.

Klašnja, Marko, and Rocío Titiunik. 2013. "Incumbency Disadvantage in Weak Party Systems: Evidence from Brazil." Unpublished manuscript, University of Michigan.

Lee, David S. 2008. "Randomized Experiments from Non-Random Selection in U.S. House Elections." *Journal of Econometrics* 142(2): 675–97.

Lee, David S., Enrico Moretti, and Matthew J. Butler. 2004. "Do Voters Affect or Elect Policies? Evidence from the U.S. House." *Quarterly Journal of Economics* 119(3): 807–59.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature* 48(2): 281–355.

Lehoucq, Fabrice. 2003. "Electoral Fraud: Causes, Types, and Consequences." *Annual Review of Political Science* 6(1): 233–56.

Leigh, Andrew. 2008. "Estimating the Impact of Gubernatorial Partisanship on Policy Settings and Economic Outcomes: A Regression Discontinuity Approach." *European Journal of Political Economy* 24(1): 256–68.

Linden, Leigh L. 2004. "Are Incumbents Really Advantaged? The Preference for Non-Incumbents in Indian National Elections." Unpublished manuscript, University of Texas at Austin.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics* 142(2): 698–714.

Miratrix, Luke W., Jasjeet S. Sekhon, and Bin Yu. 2013. "Adjusting Treatment Effect Estimates by Post-Stratification in *Randomized Experiments*." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 75(2): 369–96.

Pettersson-Lidbom, Per. 2008. "Do Parties Matter for Economic Outcomes? A Regression-Discontinuity Approach." *Journal of the European Economic Association* 6(5): 1037–56.

Querubin, Pablo. 2011. "Political Reform and Elite Persistence: Term Limits and Political Dynasties in the Philippines." Unpublished manuscript, New York University.

Rubin, Donald B. 1973. "The Use of Matched Sampling and Regression Adjustment to Remove Bias in Observational Studies." *Biometrics* 29(1): 185–203.

Rubin, Donald B. 1979. "Using Multivariate Matched Sampling and Regression Adjustment to Control Bias in Observational Studies." *Journal of the American Statistical Association* 74(366a): 318–28.

Rubin, Donald B. 2009. "Should Observational Studies Be Designed to Allow Lack of Balance in Covariate Distributions across Treatment Groups?" *Statistics in Medicine* 28(9): 1420–23.

Schochet, Peter Z. 2010. "Is Regression Adjustment Supported by the Neyman Model for Causal Inference?" *Journal of Statistical Planning and Inference* 140(1): 246–59.

Simpser, Alberto. 2013. *Why Governments and Parties Manipulate Elections: Theory, Practice, and Implications.* New York: Cambridge University Press.

Snyder, Jason. 2005. "Detecting Manipulation in U.S. House Elections." Unpublished manuscript, UCLA.

Thistlethwaite, Donald L., and Donald T. Campbell. 1960. "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment." *Journal of Educational Psychology* 51(6): 309–17.

Trounstine, Jessica. 2011. "Evidence of a Local Incumbency Advantage." *Legislative Studies Quarterly* 36(2): 255–80.

Uppal, Yogesh. 2009. "The Disadvantaged Incumbents: Estimating Incumbency Effects in Indian State Legislatures." *Public Choice* 138(1–2): 9–27.

Uppal, Yogesh. 2010. "Estimating Incumbency Effects in U.S. State Legislatures: A Quasi-Experimental Study." *Economics & Politics* 22(2): 180–99.

Vogl, Tom. 2012. "Race and the Politics of Close Elections." Working paper, Princeton University.

Westfall, Peter H., and Stanley S. Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment.* New York: Wiley.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's website:

Appendix A: Data sources and definitions.
Appendix B: Graphs.
Appendix C: Model of pre-electoral manipulation.