Misunderstandings About the Regression Discontinuity Design in the Study of Close Elections*

Brandon de la Cuesta[1] and Kosuke Imai[2]

[1]Department of Politics, Princeton University, Princeton, New Jersey 08544; email: bjmiller@princeton.edu

[2]Department of Politics and Center for Statistics and Machine Learning, Princeton University, Princeton, New Jersey 08544; email: kimai@princeton.edu

Annu. Rev. Polit. Sci. 2016. 19:375–96

First published online as a Review in Advance on March 2, 2016

The *Annual Review of Political Science* is online at polisci.annualreviews.org

This article's doi:
10.1146/annurev-polisci-032015-010115

Keywords

as-if-random assumption, continuity, extrapolation, multiple testing, placebo test, sorting

Abstract

Recently, the regression discontinuity (RD) design has become increasingly popular among social scientists. One prominent application is the study of close elections. We explicate several methodological misunderstandings widespread across disciplines by revisiting the controversy concerning the validity of RD design when applied to close elections. Although many researchers invoke the local or as-if-random assumption near the threshold, it is more stringent than the required continuity assumption. We show that this seemingly subtle point determines the appropriateness of various statistical methods and changes our understanding of how sorting invalidates the design. When multiple-testing problems are also addressed, we find that evidence for sorting in US House elections is substantially weaker and highly dependent on estimation methods. Finally, we caution that despite the temptation to improve the external validity, the extrapolation of RD estimates away from the threshold sacrifices the design's advantage in internal validity.

*A replication archive for all figures and reported results is available as de la Cuesta, Brandon; Imai, Kosuke, 2016, "Replication Data for: Misunderstandings about the Regression Discontinuity Design in Close Elections Dataverse," **http://dx.doi.org/10.7910/DVN/HWH5YN**, Harvard Dataverse.

INTRODUCTION

During the past decade, the regression discontinuity (RD) design, first developed by education policy researchers more than half a century ago (Thistlewaite & Campbell 1960), has become increasingly popular among social scientists. This is in part due to the fact that the RD design can provide valid causal estimates under relatively weak assumptions in observational studies (see Imbens & Lemieux 2008, Lee & Lemieux 2010, Skovron & Titiunik 2015, for useful review articles). The validity of the RD design does not require the presence of randomized treatments. Instead, researchers must find a forcing variable that deterministically assigns treatments to units based on whether their values of the forcing variable are above or below a known threshold. Although various applications of the RD design exist in the social sciences, the most prominent application in political science has been the study of close elections. **Table 1** presents a list of articles published in three major political science journals over the last decade that utilize the RD design. The table shows that the use of close elections has been by far the most frequent application of this design in political science.[1]

Despite its popularity, however, there exists a considerable debate in the literature as to the validity of applying the RD design to close elections. In their influential study, Lee et al. (2004) present evidence supporting the credibility of the RD design in the case of US House of Representatives elections (see also Lee 2008). Recent studies cast significant doubt on this conclusion (Caughey & Sekhon 2011, Grimmer et al. 2011, Snyder 2005), while others defend the applicability of the RD design (Eggers et al. 2015a). In this article, we revisit this controversy. We demonstrate that once all important methodological issues are properly addressed, the RD design appears to be valid in the case of US House elections. Given our focus on the applications to close elections, however, our discussion is confined to the sharp RD design, in which the forcing variable completely determines the treatment assignment. We do not discuss the fuzzy RD design, in which the treatment assignment is not a deterministic function of the forcing variable.

While addressing this controversy, we also seek to clarify several methodological misunderstandings about the RD design. To do so, we divide the article into three parts. First, we discuss

Table 1 Recently published applications of the regression discontinuity (RD) design in three major political science journals

Journal[a]	Studies applying RD to close elections	Studies applying RD to other topics
APSR	Eggers & Hainmueller (2009) Galasso & Nannicini (2011)	Posner (2004), geography Dunning & Nilekani (2013), ethnic population Samii (2013), retirement age
AJPS	Gerber & Hopkins (2011) Folke & Snyder (2012) Eggers et al. (2015a)	Dinas (2014), age Holbein & Hillygus (2016), age
JOP	Gerber et al. (2011) Boas et al. (2014) Fouirnaies & Hall (2014) Erikson et al. (2015) Hainmueller et al. (2015)	Krasno & Green (2008), geography Friedman & Holden (2009), geography

[a] APSR, *American Political Science Review*; AJPS, *American Journal of Political Science*; JOP, *Journal of Politics*.

[1] See Eggers et al. (2015b) for a review of studies that exploit population thresholds of municipalities.

the identification problem under the RD design. We point out that although many researchers invoke the local randomization assumption, also called the as-if-random assumption, it tends to be more stringent than the continuity assumption, which is the key identification assumption of the RD design. The local randomization assumption states that within a window of prespecified size around the discontinuity threshold, whether or not an observation receives the treatment is essentially randomly determined. This assumption implies that observations on one side of the threshold are on average identical to those on the other side of it in terms of any pretreatment covariates. In contrast, the continuity assumption requires that the only change, which occurs at the point of discontinuity, is the shift in the treatment status. Under the continuity assumption, observations on either side of the discontinuity threshold can systematically differ from each other in many aspects, even by a large magnitude. We demonstrate that this seemingly subtle difference between the two assumptions can alter the understanding of how sorting invalidates the RD design.

Next, we turn to the issue of estimation and inference under the RD design. We show that the two assumptions—the local randomization assumption and the continuity assumption—lead to a divergent choice of estimation methods and can consequently alter empirical findings. We present empirical evidence that the use of the difference-in-means estimator, which is based on the local randomization assumption, is particularly ill-suited to the close elections application. Although a better approach is to fit a linear regression on each side of the discontinuity threshold, this still requires researchers to specify the size of the window around the threshold. To avoid the arbitrariness of this choice and enable more flexible modeling, researchers can use a local linear regression combined with an optimal, data-driven bandwidth selection procedure developed in the literature (Imbens & Kalyanaraman 2012, Calonico et al. 2014). This method is known to have better theoretical properties at the discontinuity threshold (Fan & Gijbels 1996). The idea is to fit a weighted linear regression on either side of the threshold, with observations farther away from the threshold assigned smaller weights. The local linear regression, therefore, offers flexibility with little loss of statistical power while removing an arbitrary choice of window size from researchers. On these grounds, we recommend that the local linear estimator should be the method of choice for RD-based analysis.

Valid inference, however, requires more than a well-suited estimator. To this end, we also advocate the use of multiple-testing correction to reduce the chance of falsely concluding that the RD design is invalid. Typically, researchers conduct placebo tests in order to examine whether the key identification assumption of the RD design is credible in a particular application. These placebo tests often involve the examination of evidence for discontinuities in a large number of pretreatment covariates, leading to many statistical tests. It is well known, however, that conducting many statistical tests can result in false rejection of null hypotheses even when all the null hypotheses under consideration are valid. When multiple-testing problems are addressed, we find that evidence for sorting in US House elections is substantially weaker and highly dependent on estimation methods.

Finally, we discuss the external validity of the RD design. The strong internal validity of this design comes with poor external validity. In the study of close elections, only the party incumbency advantage in elections with exact ties is identifiable. In order to overcome this major limitation, researchers have attempted to extrapolate the RD estimates away from the threshold (Hainmueller et al. 2015) using the method recently proposed by Angrist & Rokkanen (2015). We caution that this approach requires researchers to rely on a version of the local randomization assumption, thereby sacrificing the internal validity of the RD design. In fact, no approach, including this methodology, is able to overcome the fact that the observed data are completely uninformative about how the extrapolation should be performed.

IS THE REGRESSION DISCONTINUITY DESIGN APPLICABLE TO CLOSE ELECTIONS?

The RD design was first applied to close elections by Lee et al. (2004). The authors utilized the Democratic vote margin to examine the effect of close elections on politicians' subsequent roll-call voting behavior. Lee et al. (2004) report the results of several placebo tests using the Democratic margin as the forcing variable within windows of various size above and below the threshold. Their covariates include the size of the African-American electorate, urbanization, and high school graduation rates. Finding no statistically significant average differences between the observations above and below the threshold within those windows, Lee et al. (2004, p. 837) conclude:

> Overall, the evidence strongly supports a valid regression discontinuity design. And as a consequence, it appears that among close elections, who wins appears virtually randomly assigned, which is the identifying assumption of our empirical strategy.

In a subsequent paper, Lee (2008) uses the same RD design to estimate the party incumbency advantage, i.e., the causal effect of a party winning the current election on its vote share in the next election. The author tests the continuity of pretreatment covariates using a range of parametric specifications, relying primarily on a fourth-order polynomial regression. Lee finds no systematic evidence of discontinuity in the Democratic vote margin of the previous election. This variable constitutes a key confounder because it is strongly correlated with both the outcome variable (the outcome of the next election) and treatment variable (the outcome of the current election). Based on this result, Lee concludes that the RD design is applicable to close US House elections.

The argument for the validity of the RD design in close elections is further bolstered by McCrary (2008), who proposes an estimator designed to test the continuity of the density function of the forcing variable. The intuition behind McCrary's approach is straightforward. If agents are able to sort themselves across a given threshold, we should expect the proportion of observations just to the left of the cutpoint to be substantially different from those to the right. Sorting, if it exists, would therefore produce a discontinuity not only in the distribution of pretreatment covariates but also in the density of the forcing variable. In the context of close elections, we would expect sorting to yield a larger number of close elections in which the Democratic candidates barely win than those in which they barely lose. Nevertheless, McCrary finds little evidence of discontinuity in the density function of the Democratic margin and concludes that there is no indication of sorting in the US House elections with respect to the density of the forcing variable.

Claiming to have discovered new evidence in favor of sorting, Caughey & Sekhon (2011) challenged the validity of the RD design in the US House elections. The authors found errors in the original dataset used by Lee and created a new dataset with several additional covariates to revisit the findings of Lee (2008) and McCrary (2008). Caughey & Sekhon (2011) found empirical evidence of sorting in close elections, even in a narrow window of one half percentage point from the threshold. They contend that because the most imbalanced covariates are related to the incumbents and their resources, sorting is most likely attributable to a general ability of well-organized, well-financed campaigns to win close elections by influencing vote totals on or before Election Day. As an example, they argue that such campaigns are able to monitor and, when necessary, intervene in vote tallies on Election Day and to convince sympathetic judges to extend polling hours in friendly precincts (Caughey & Sekhon 2011, p. 397).

The controversial findings of Caughey & Sekhon rely on two new analyses that involve additional covariates and particular subsetting. The authors conduct placebo tests utilizing the incumbency status in the previous election and several other substantively meaningful covariates.

Examining the relationship between the forcing variable and these additional covariates near the threshold, they find larger covariate imbalance in narrower windows around the threshold. They show that the outcomes of elections even in the one-half-percentage-point window appear less random than those farther away. Specifically, the authors conduct two nonparametric tests and find the covariate imbalance to be largest for measures of previous political experience, incumbency, pre–Election Day donations, and total campaign spending, precisely the variables that would appear to most greatly affect a candidate's ability to win a very close election. As additional evidence, Caughey & Sekhon (2011) cite the ability of *Congressional Quarterly* to correctly predict 31 of 44 very close races. They conclude: "Far from being randomly decided, the outcomes of very close elections are actually quite predictable" (p. 393).

Caughey & Sekhon (2011) conducted a second analysis to question the validity of the RD design by subsetting the data with incumbency status. Performing the density test proposed by McCrary (2008) separately for districts where the seat was previously held by a Democrat and those where the Democratic candidate was a challenger, they show that the pooled test run by McCrary may have masked important variation by incumbency status. The density of the forcing variable on the incumbent-only sample appears to be highly imbalanced at the cutpoint, with Democratic candidates more likely to win close elections in seats they won in the previous election. Caughey & Sekhon (2011) argue that because such imbalance exists for both Democratic and Republican incumbents, the pooled analysis of McCrary (2008) failed to identify this discontinuity.

A notable exception to the pattern of the results reported by Caughey & Sekhon (2011) is the lack of significant results for two measures: the party of the governor and the party of the secretary of state. These measures are expected to be closely correlated with a party's ability to influence, legally or otherwise, close elections. In an unpublished manuscript, Grimmer et al. (2011) analyze a new dataset that includes all US House elections from 1880 to 2008 and conduct placebo tests for these measures by fitting a third-order polynomial regression within a 10-percentage-point window around the threshold. Their analysis suggests that candidates of parties who control the governorship, secretary of state office, or state legislature are substantially more likely to win close elections than candidates of out-parties, often by several percentage points. Together, the evidence for sorting presented by Caughey & Sekhon (2011) and Grimmer et al. (2011) appears to invalidate the application of the RD design to close elections. This argument is bolstered by Snyder (2005), who demonstrates the disproportionate rate with which incumbents win close elections.

A recent study by Eggers et al. (2015a) brings additional quantitative and primary source data to bear on the question of sorting. Analyzing data from more than 40,000 close races across several countries, Eggers and colleagues argue that existing evidence for the US House case is based in part on inappropriate tests. The authors also show that the imbalance in pretreatment covariates in the post–World War II dataset does not exist in a larger dataset going back to 1880 or in other advanced, industrialized democracies.[2] In particular, the authors suggest that the use of the difference-in-means estimator has likely led to biased inference, a point we also make below, due to the strong correlation between pretreatment covariates and the forcing variable even in very narrow windows. They also find no evidence of discontinuities in the density of the running variable according to incumbency status, contradicting another key finding from Caughey & Sekhon (2011). Finally, Eggers et al. (2015a) provide an informative discussion of the substantive plausibility of various sorting mechanisms. They find that resource-type explanations require a

[2] In addition to expanding the number of countries, Eggers et al. (2015a) also look for discontinuities in post–World War II statewide office (since 1947), state legislature (since 1990), and mayoral (since 1947) races.

much greater degree of manipulation and information on expected vote share than even modern campaigns appear to possess.

In a detailed analysis of the observed covariate imbalance in the Caughey & Sekhon (2011) dataset, Erikson & Rader (2013) make a similar argument. They demonstrate that other variables that might plausibly measure the source of incumbent advantages in close elections are balanced in the one-half-percentage-point window. The authors analyze the number of cases that are responsible for the relatively larger share of incumbents on the right-hand side of the cutpoint—that is, those who won elections. They find that, even if the observed covariate imbalance originates from actual advantage in winning close elections, its impact on the resulting estimated increase in $t + 1$ vote share would be minimal, perhaps as low as 0.17 percentage points. Erikson & Rader (2013) thus suggest that, regardless of potential violations of continuity, the resulting bias is likely to be small.

With some authors reporting strong evidence against the validity of the RD design in US House elections and others supporting its use, the literature is remarkably divided on the question of whether sorting exists in the close election context. As Eggers et al. (2015a) note, this is in part due to differences in the time period under consideration. It is also, though perhaps to a lesser degree, due to differences in the covariates used by researchers on either side of the debate. Yet, just as the covariates used to evaluate the continuity assumption differ, so too does the method by which those discontinuities are estimated. Surveying the major articles about the validity of the RD design in close elections, we count no fewer than six unique combinations of method and window selection preferred by different authors (see also Skovron & Titiunik 2015, who report similar findings in other applications of the RD design).

In the remainder of this article, we shed light on how different methodologies can yield substantively divergent conclusions regarding the validity of the RD design when applied to close elections. We show that these methodological issues are directly related to the RD design's continuity assumption and therefore are also relevant in other applications.

THE CONTINUITY ASSUMPTION DOES NOT IMPLY THE LOCAL RANDOMIZATION ASSUMPTION

In this section, we first clarify a common misunderstanding of the key identification assumption under the RD design. In the literature, researchers often invoke the local randomization or as-if-random assumption. In a review, Lee & Lemieux (2010, p. 283) effectively summarize this position:

> RD designs can be analyzed—and tested—like randomized experiments. This is the key implication of the local randomization result. If variation in the treatment near the threshold is approximately randomized, then it follows that all "baseline characteristics"—all those variables determined prior to the realization of the assignment variable—should have the same distribution just above and just below the cutoff.

This local randomization assumption, therefore, implies that in close elections—within some prespecified window near the 50–50 margin—whether a candidate becomes a barely-winner or a barely-loser is randomly determined. The local randomization assumption is often motivated by focusing on the way in which imprecise manipulation of the forcing variable will produce a local randomized experiment near the cutpoint. Lee & Lemieux (2010) suggest that inability to sort across the threshold, and the random variation in treatment assignment it induces, is required for the RD design to be valid. Other scholars have made a similar argument, justifying the validity of the RD design on the grounds of local randomization. For example, Dunning (2008, p. 289), discussing the advantages and disadvantages of natural experiments, writes that "the claim of 'as if' random assignment in the neighborhood of the threshold may be especially plausible in regression-discontinuity designs" (see Samii 2013 for another example).

However, the local randomization assumption is not required for the RD design, a point also made explicitly by Cattaneo et al. (2015a) and Skovron & Titiunik (2015). Indeed, this assumption—which will hold if the treatment is randomly assigned near the threshold—is typically more stringent than the continuity of expected potential outcomes, which is the key identification assumption of the RD design.[3] The continuity assumption does not necessarily imply that the treatment assignment is randomly determined in a narrow window near the threshold. Thus, the finding that barely-winners and barely-losers are different in a number of dimensions near the threshold does not necessarily invalidate the application of the RD design. Indeed, the continuity assumption only requires that the sole change occurring at the discontinuity point is the shift in the treatment status.[4]

To see this seemingly subtle difference between the two assumptions more formally, let $Y_i(1)$ and $Y_i(0)$ represent the potential outcomes for unit i under the treatment ($T_i = 1$) and control ($T_i = 0$) conditions, respectively. In the study of close elections, $Y_i(1)$ ($Y_i(0)$) represents the outcome variable of interest for district i, e.g., the Democratic vote share in the next election, under the scenario that a Democratic (Republican) candidate wins the election in the current period, $T_i = 1$ ($T_i = 0$). Thus, in this context, the Democratic victory is the treatment condition and the Republican victory is the control condition.

We use X_i to denote the forcing variable for unit i. In the close election context, X_i represents the Democratic vote margin for district i. It is important to note that the forcing variable is *deterministically* related to the treatment assignment. Specifically, the unit is assigned to the control condition if X_i is below the threshold c and to the treatment condition if X_i exceeds c, i.e., $T_i = \mathbf{1}\{X_i > c\}$. Therefore, the treatment assignment probability given X_i is either 0 or 1. This contrasts with the fact that the treatment assignments are assumed to be *stochastically* determined under the local randomization assumption.

Under this setup, the continuity assumption can be formally written as

$$\mathbb{E}(Y_i(1)|X_i = c) = \lim_{x \downarrow c} \mathbb{E}(Y_i(1)|X_i = x), \qquad 1.$$

$$\mathbb{E}(Y_i(0)|X_i = c) = \lim_{x \uparrow c} \mathbb{E}(Y_i(0)|X_i = x), \qquad 2.$$

where the limit is taken from above the threshold in the first equation and from below it in the second. The equations imply that there is no discontinuous jump in the conditional expectation

[3] This statement assumes exclusion restriction as well as local randomization. It is possible that the forcing variable can be assumed to be randomized near the threshold but still influence the outcome in such a way that the continuity assumption is violated. In the context of close elections, even if the margin of victory can be assumed to be randomized within a window around the threshold, the vote share of the previous election may directly affect the outcome of the next election in some other way than through the binary treatment variable. Our subsequent discussion ignores this possible violation of the exclusion restriction. See Cattaneo et al. (2015a) for a more detailed discussion on this point.

[4] Although the local randomization assumption is not a requirement for RD estimates to be valid, Cattaneo et al. (2015a) propose an objective criterion for establishing the window in which randomization appears to hold and propose the use of randomization inference to estimate treatment effects under the local randomization assumption. In subsequent work, Skovron & Titiunik (2015) review this approach and provide a comprehensive discussion of the local randomization framework as well as note its substantial differences from the continuity-based framework we focus on here. See also Cattaneo et al. (2015c) for an application of the randomization inference framework on the Head Start program.

function of each potential outcome.[5] Under this assumption, therefore, we can identify the average treatment effect at the threshold c,

$$\begin{aligned}\mathbb{E}(Y_i(1) - Y_i(0)|X_i = c) &= \lim_{x \downarrow c} \mathbb{E}(Y_i(1)|X_i = x) - \lim_{x \uparrow c} \mathbb{E}(Y_i(0)|X_i = x) \\ &= \lim_{x \downarrow c} \mathbb{E}(Y_i|X_i = x) - \lim_{x \uparrow c} \mathbb{E}(Y_i|X_i = x). \qquad 3.\end{aligned}$$

In contrast, the local randomization assumption demands that within a prespecified window $[c_0, c_1]$ where $c_0 < c < c_1$, the treatment assignment is randomized,

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp 1\{X_i > c\}|c_0 \leq X_i \leq c_1, \qquad 4.$$

where $\perp\!\!\!\perp$ denotes statistical independence. The assumption implies that, within the window, the average potential outcomes are identical below and above the threshold:

$$\mathbb{E}(Y_i(t)|c_0 \leq X_i \leq c) = \mathbb{E}(Y_i(t)|c < X_i \leq c_1) \qquad 5.$$

for $t = 0, 1$. Formally, this follows from the fact that Equation 4 implies $\mathbb{E}(Y_i(t)|X_i \leq c, c_0 \leq X_i \leq c_1) = \mathbb{E}(Y_i(t)|X_i > c, c_0 \leq X_i \leq c_1)$.

Whereas the continuity assumption will hold whenever there is local randomization,[6] the opposite is not the case. In fact, nothing in the continuity assumption requires the expected potential outcomes on both sides of the threshold to be identical. This difference between the two assumptions becomes critical when the forcing variable, which completely determines the treatment assignment, is related to the potential outcomes because it acts as a strong confounder. Estimators designed to test the distributional equivalence implied by the local randomization assumption may thus falsely discover discontinuities even in cases where the continuity assumption holds.[7]

We graphically illustrate why the local randomization assumption is stronger than the continuity assumption. **Figure 1** shows two empirical examples of pretreatment covariates that exhibit a strong (**Figure 1*a***, based on the Democratic experience advantage) or moderate (**Figure 1*b***, based on the proportion of total election spending) relationship with the forcing variable. Under the local randomization assumption, the estimated discontinuity is based on two flat lines with zero slope (red dashed lines) because the treatment and control groups are assumed to be identical on average within a prespecified threshold—in this case, $[-0.02, 0.02]$, indicated by dotted vertical lines. In contrast, under the continuity assumption, we do not assume the absence of the association between the outcome and forcing variables (blue solid lines). The figure illustrates that the local randomization assumption can falsely discover a discontinuity (**Figure 1*a***) or overestimate one (**Figure 1*b***).

The figure also demonstrates a key feature of the close elections data, namely that many pretreatment covariates exhibit a strong relationship with the forcing variable near the cutpoint. That this relationship exists so frequently is to be expected in the electoral context, where sophisticated actors compete in a zero-sum game in which candidate quality and other related characteristics affect the outcome indirectly through the forcing variable. Wherever this is the case, one must model this relationship in a principled and transparent way.

[5] Technically, although we call this the continuity assumption, we are only assuming that the conditional expectation of potential outcome at the threshold can be approximated well from one side rather than from both sides.

[6] As explained in footnote 3, this statement assumes the exclusion restriction that the conditional expectation function does not depend on the value of the forcing variable within the window.

[7] Imbens & Lemieux (2008) make a similar point, also echoed by Eggers et al. (2015a), that this bias is likely to be high in cases with strong correlation between the forcing variable and the outcome of interest.

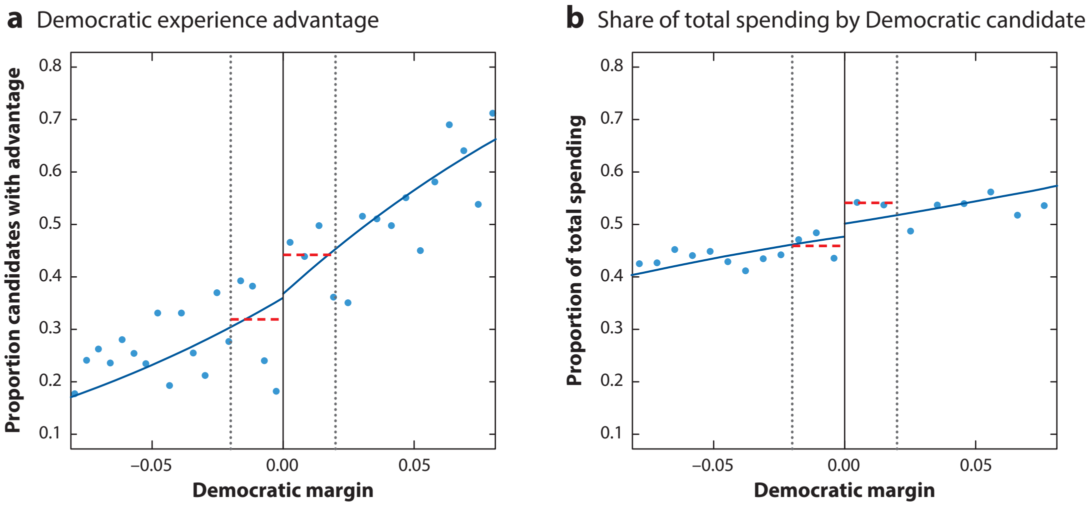

Figure 1

The problem of the local randomization assumption. Under the local randomization assumption, also called the as-if-random assumption, the observations below and above the discontinuity threshold, a [−0.02, 0.02] window indicated by dotted lines in this case, are assumed to be identical on average. As a result, the estimated discontinuity is based on two flat lines with no slope (red dashed lines). In contrast, under the continuity assumption, the association with the forcing variable is not assumed to be absent (blue solid lines). The two plots are based on the dataset on US House elections by Caughey & Sekhon (2011) using two pretreatment covariates: the experience advantage of the Democratic candidate (*a*) and the proportion of total donations given to the Democrat (*b*). They show that the local randomization assumption can falsely discover a discontinuity (*a*) or overestimate one (*b*).

WHEN IS THE CONTINUITY ASSUMPTION VIOLATED?

The discussion so far implies that imbalance in pretreatment covariates just below and above the threshold does not necessarily imply the violation of the identification assumption for the RD design. Under the continuity assumption, such imbalance can exist so long as there is no discontinuous jump at the threshold. Lack of discontinuity in pretreatment covariates at the threshold then represents empirical evidence for the continuity of the expected potential outcomes so long as all pretreatment covariates relevant for the outcome of interest are measured and analyzed.

If covariate imbalance does not necessarily invalidate the RD design, what are the scenarios under which the continuity assumption is violated? To answer this question, we must consider the kind of sorting behavior that pushes would-be barely-losers up above the threshold or moves potential barely-winners down below the threshold.[8] Such sorting will lead to a discontinuous jump at the threshold in the conditional expectation function of the potential outcomes. This in turn is likely to manifest as a discontinuous jump in pretreatment covariates, which are associated with the outcome.

Following the informative discussion of this issue by Eggers et al. (2015a), we consider two types of sorting behavior. One is due to pre-election behavior or characteristics of candidates, whereas the other is owing to postelection advantages in vote tallying, including the ability to engineer electoral fraud. We argue that although the postelection sorting behavior clearly constitutes a violation of the continuity assumption, the pre-election behavior may not. The occurrence of electoral fraud, for example, implies that a candidate who would have barely lost the election ends

[8] See Eggers et al. (2015b) for a discussion of sorting behavior when using population thresholds as the discontinuity cutoffs.

up becoming a winner. In other words, the election fraud pushes above the winning threshold the candidate observations that would have been located just below it, creating a discontinuous jump at the threshold in the conditional expectation function of the potential outcomes.[9]

Next, consider the potential sorting based on pre-election candidate behavior and characteristics. Although they do not advocate it themselves, Grimmer et al. (2011, p. 10) describe this hypothesis as follows: "When campaigns know that an election will be close (either through partisan information networks or polls), they invest more resources and effort in those contests." Caughey & Sekhon (2011) argue that in a close election, incumbent campaigns engender heavy investment of financial and institutional resources to win the race. Yet, if the argument is that in close elections, the intensity and amount of investment, whether in the form of advertising or get-out-the-vote mobilization, are high, then the continuity assumption will not be violated. The increased deployment of resources in close elections is likely to strengthen the association between the expected potential outcomes and the forcing variable rather than to produce a discontinuous relationship at the threshold. For example, the candidates who manage to attract abundant campaign contributions in the current election may be likely to receive a higher vote share in the next election as well. As Eggers et al. (2015a) explain, however, in order for the pre-election sorting behavior to exist, campaigns would need to be able to predict vote shares with extreme precision (e.g., a quarter of one percentage point). The qualitative evidence presented by the authors suggests such a scenario is highly unlikely.

In summary, the violation of the continuity assumption calls for a substantive scenario about particular sorting behavior that results in a discontinuous jump at the threshold. Postelection sorting behavior such as election fraud may imply the violation of the continuity assumption because it would put some potential losers just above the winning threshold. However, pre-election sorting behavior would require the campaign of the eventual winner to be able to predict election outcomes with extreme accuracy, and then deploy necessary resources to win the race. Existing evidence suggests that the latter scenario is unlikely. Furthermore, the close elections case is also instructive for researchers seeking to apply the RD design in other contexts because it demonstrates that a nuanced understanding of the continuity assumption will provide guidance on the substantive form that sorting may take in specific applications.

METHODOLOGICAL CHOICE FOR TESTING THE VALIDITY OF THE REGRESSION DISCONTINUITY DESIGN

The crucial difference between the local randomization and continuity assumptions also affects the choice of methods for testing the validity of the RD design and may alter empirical conclusions. As illustrated by the Lee & Lemieux (2010) quotation that began the third section of this article, under the local randomization assumption, researchers look for evidence of imbalance in pretreatment covariates within a selected window near the threshold. The simplest way of doing so is to use the difference-in-means estimator, applied to a pretreatment covariate Z using the observations in the window, $X_i \in [c_0, c_1]$. This estimator is formally defined as follows:

$$\hat{\tau}_{\mathrm{DM}}(Z; X, c_0, c_1) = \frac{1}{n_{0c}} \sum_{i=1}^{n} 1\{c_0 \leq X_i \leq c\} Z_i - \frac{1}{n_{1c}} \sum_{i=1}^{n} 1\{c < X_i \leq c_1\} Z_i, \qquad 6.$$

[9] This argument holds unless randomly selected candidates are assumed to commit election fraud under the local randomization assumption.

where n_{0c} (n_{1c}) is the total number of observations below (above) the threshold c in the window, i.e., $n_{0c} = \sum_{i=1}^{n} \mathbf{1}\{c_0 \leq X_i \leq c\}$ and $n_{1c} = \sum_{i=1}^{n} \mathbf{1}\{c < X_i \leq c_1\}$. As done by Caughey & Sekhon (2011), one can also compare the distributions, rather than their means, between the two groups of observations within this window by using popular nonparametric tests such as the Fisher's exact test, the Kolmogorov–Smirnov (KS) test, and the Wilcoxon's rank-sum test. Regardless of which test is used, researchers rely on the as-if-random assumption that the covariate distribution should be similar between the treated and control groups near the threshold.

As explained earlier, the local randomization assumption tends to be more stringent than the continuity assumption required for the RD design. As a result, when conducting placebo tests, we must account for possible association between the pretreatment covariate Z and the forcing variable X. We can regress Z on X within the preselected window near the threshold, $[c_0, c_1]$. This linear regression estimator is defined as the difference between the estimated intercepts from the two regressions,

$$\hat{\tau}_{\mathrm{LR}}(Z; X, c_0, c_1) = \hat{\alpha}_1 - \hat{\alpha}_0, \qquad 7.$$

where

$$(\hat{\alpha}_0, \hat{\beta}_0) = \underset{\alpha_0, \beta_0}{\arg\min} \sum_{i=1}^{n} 1\{c_0 \leq X_i \leq c\}\{Y_i - \alpha_0 - \beta_0(X_i - c)\}^2 \qquad 8.$$

and

$$(\hat{\alpha}_1, \hat{\beta}_1) = \underset{\alpha_1, \beta_1}{\arg\min} \sum_{i=1}^{n} 1\{c < X_i \leq c_1\}\{Y_i - \alpha_1 - \beta_1(X_i - c)\}^2. \qquad 9.$$

Note that when the slope parameters, β_0 and β_1, are assumed to be exactly zero, this linear regression estimator is identical to the difference-in-means estimator. Thus, assuming the local randomization of treatment assignment leads to the restriction of zero slopes when testing the validity of the RD design. Such a restriction, if inconsistent with the data, can alter the results of empirical tests.

One drawback of the two tests described above is that the results can be sensitive to the choice of window size. In the literature on close elections, windows of about two percentage points are common (e.g., Lee 2008, Butler 2009). Caughey & Sekhon (2011) choose a more restrictive window size of plus–minus half a percentage point. There is a clear bias–variance trade-off here. If researchers choose a window that is too narrow, placebo tests will lack statistical power and yield many false negatives. That is, even if the RD design is invalid, statistical tests may fail to detect it because they are based on too few observations. However, if the window is too wide, tests may yield many false positives because their assumptions no longer hold. For example, even when the linear regression approximates the true data-generating process well in a narrow window, the approximation may become poor when the window is widened to include more observations.

Recently, new methods have been developed so that researchers can avoid this arbitrary choice of window size (Imbens & Kalyanaraman 2012, Calonico et al. 2014). These methods are based on the local linear regression, a nonparametric generalization of the linear regression estimator discussed above. These local linear regression estimators assign smaller weights to observations far from the threshold. The estimator is based on the following weighted linear regressions:

$$\tau_{\mathrm{LLR}}(Z; X, K, h) = \hat{\alpha}_1 - \hat{\alpha}_0, \qquad 10.$$

$$(\hat{\alpha}_0, \hat{\beta}_0) = \underset{\alpha_0, \beta_0}{\arg\min} \sum_{i=1}^{n} 1\{c_0 \leq X_i \leq c\}\{Y_i - \alpha_0 - \beta_0(X_i - c)\}^2 K\left(\frac{X_i - c}{h}\right), \qquad 11.$$

and

$$(\hat{\alpha}_1, \hat{\beta}_1) = \underset{\alpha_1, \beta_1}{\arg\min} \sum_{i=1}^{n} 1\{c < X_i \leq c_1\}\{Y_i - \alpha_1 - \beta_1(X_i - c)\}^2 K\left(\frac{X_i - c}{h}\right), \qquad 12.$$

where $K(\cdot)$ is the weighting or Kernel function. The choice of Kernel function and its bandwidth parameter, h, control for the weighting scheme.

To see that the estimator based on linear regression, τ_{LR}, is a special case of that based on local linear regression, $\tau_{\text{LLR}}(Z; X, K, h)$, consider the uniform Kernel, recommended by Imbens & Kalyanaraman (2012), which is formally defined as

$$K(u) = \frac{1}{2} \cdot \mathbf{1}\{|u| < 1\}. \qquad 13.$$

This Kernel gives an equal positive weight to each observation within the window $[-h + c, c + h]$ while giving zero weight to those outside it. Thus, this Kernel function gives an estimator identical to the one based on linear regression in the window when $c_0 = c - h$ and $c_1 = c + h$. Another popular Kernel function used in the literature is the triangular Kernel (Calonico et al. 2014),

$$K(u) = (1 - |u|) \cdot \mathbf{1}\{|u| < 1\}, \qquad 14.$$

which assigns smaller weights to observations far from the threshold within the window $[-h + c, c + h]$.

The local linear regression approach has two main advantages. First, it has been established that the local linear regression estimators have better theoretical properties at the boundary when compared to other popular parametric (e.g., regression with higher-order polynomials) and nonparametric (e.g., Kernel regression) approaches (Fan & Gijbels 1996). This is an important consideration given our goal to estimate the expected values of potential outcomes at the discontinuity threshold. Second, there exist principled, data-driven ways of selecting the weighting scheme in the methodological literature, thereby avoiding the arbitrary selection of window size. Imbens & Kalyanaraman (2012) derive the bandwith selection procedure that minimizes the approximate mean squared error at the threshold. Calonico et al. (2014) improve this approach and show how to construct a bias-corrected estimator with robust confidence intervals.[10] For this reason, we use the approach of Calonico et al. for all local linear regression results reported in this article.

In summary, we recommend that researchers use the approach based on the local linear regression with an optimal, data-driven bandwidth selection procedure developed in the literature. Unlike the difference-in-means estimator, the local linear regression can accommodate an arbitrary association between the forcing and outcome variables near the threshold. As demonstrated by **Figure 1**, the difference-in-means estimator justified under the restrictive local randomization assumption, which constrains this association to be zero, can result in a biased estimate of discontinuity. The local linear estimator is thus a more faithful test of the continuity assumption. And unlike the linear regression in the preselected window, local linear regression is also nonparametric and provides a principled way of choosing a bandwidth parameter.

[10] The confidence intervals proposed by Calonico et al. (2014) have better coverage. Their method also allows for mean-square optimal bandwidth selectors that often produce smaller bandwidths, mitigating the possibility raised by Caughey & Sekhon (2011) that the use of large bandwidths may mask variation near the cutpoint. The bandwidth selection procedure by Calonico et al. (2014) also has the added advantage that the optimal window may vary with the covariate being tested. In the results presented in the following section, for example, optimal bandwidths range from as little as 5% to as much as 28%.

EMPIRICAL CONSEQUENCES OF THE METHODOLOGICAL CHOICE

We illustrate the empirical consequences of the methodological choice discussed in the previous section by analyzing the close elections data of Caughey & Sekhon (2011). The data contain all House elections from 1946 to 2008. Following the original analysis, we use only the so-called stable districts with no missing values for the Democratic margin at times $t - 1$, t, and $t + 1$. The quantity of interest is the party incumbency advantage at the threshold, i.e., the average effect of winning a tied election on the party's vote share in the next election.

This dataset is rich and provides a marked improvement over earlier work. The additional covariates include the relative proportion of total campaign donations received by each candidate prior to the election, the share of total campaign spending, candidates' previous political experience, and whether the party held important state offices such as secretary of state and governor. In close elections, these variables are especially helpful because they measure financial and structural advantages of campaigns. For example, if we find large discontinuities in party control of state offices, which are responsible for administering elections and tabulating results, they may constitute evidence consistent with the possibility of fraud-based postelection sorting (see above, When Is the Continuity Assumption Violated?).

Following Caughey & Sekhon (2011), we analyze each of 25 pretreatment covariates to investigate the validity of the RD design within the half percentage point near the threshold. We examine whether the three methods reviewed in the previous section yield different empirical conclusions. For the difference-in-means and linear regression estimators, we use two different window sizes: $[-0.005, 0.005]$, used in the original analysis, and $[-0.02, 0.02]$, frequently employed in the close election literature. For the local linear regression, we use the method proposed by Calonico et al. (2014), available in the `R` package `rdrobust` (Calonico et al. 2015).

In this analysis, we conduct a total of 25 placebo tests using various substantively important pretreatment covariates. Conducting many placebo tests is generally a good practice, given that it is always possible that we may miss a discontinuous jump in unobserved pretreatment covariates. Nevertheless, such multiple testing creates a difficulty in interpreting the results of placebo tests. The problem is that the more statistical tests one conducts, the more likely one is to discover false positives even when all the null hypotheses are true. For example, if one conducts 25 independent placebo tests at the 0.05 level when all null hypotheses are true, the probability of falsely rejecting at least one null hypothesis exceeds 70% ($1–0.95^{25}$), with the average number of false rejections equal to 1.25. This means that we may falsely conclude that the RD design is invalid even when the continuity assumption is satisfied.[11] Throughout this article, we use the Benjamini–Hochberg procedure to control the false discovery rate, the proportion of false discoveries among discoveries (Benjamini & Hochberg 1995). Benjamini & Yekutieli (2001) show that this procedure is valid even when test statistics are positively dependent.

Figure 2 presents the results, where solid (open) circles, with solid (dashed) lines as 95% confidence intervals (not corrected for multiplicity), represent estimates in the 2-percentage-point $[-0.02, 0.02]$ (one-half-percentage-point $[-0.005, 0.005]$) window. We standardize all nonbinary variables and present estimates in terms of standard deviation units to facilitate comparison across variables. Across the three methods and different window sizes, there exists large variation in the number of pretreatment covariates for which the estimated discontinuity is statistically significant. For the difference-in-means estimator in **Figure 2*a***, 12 of the 25 covariates show a statistically significant discontinuity using a 2-percentage-point window. Of these, 5 remain significant when

[11] Caughey & Sekhon (2011) are aware of this problem and present multiple-testing corrected results in supplemental materials.

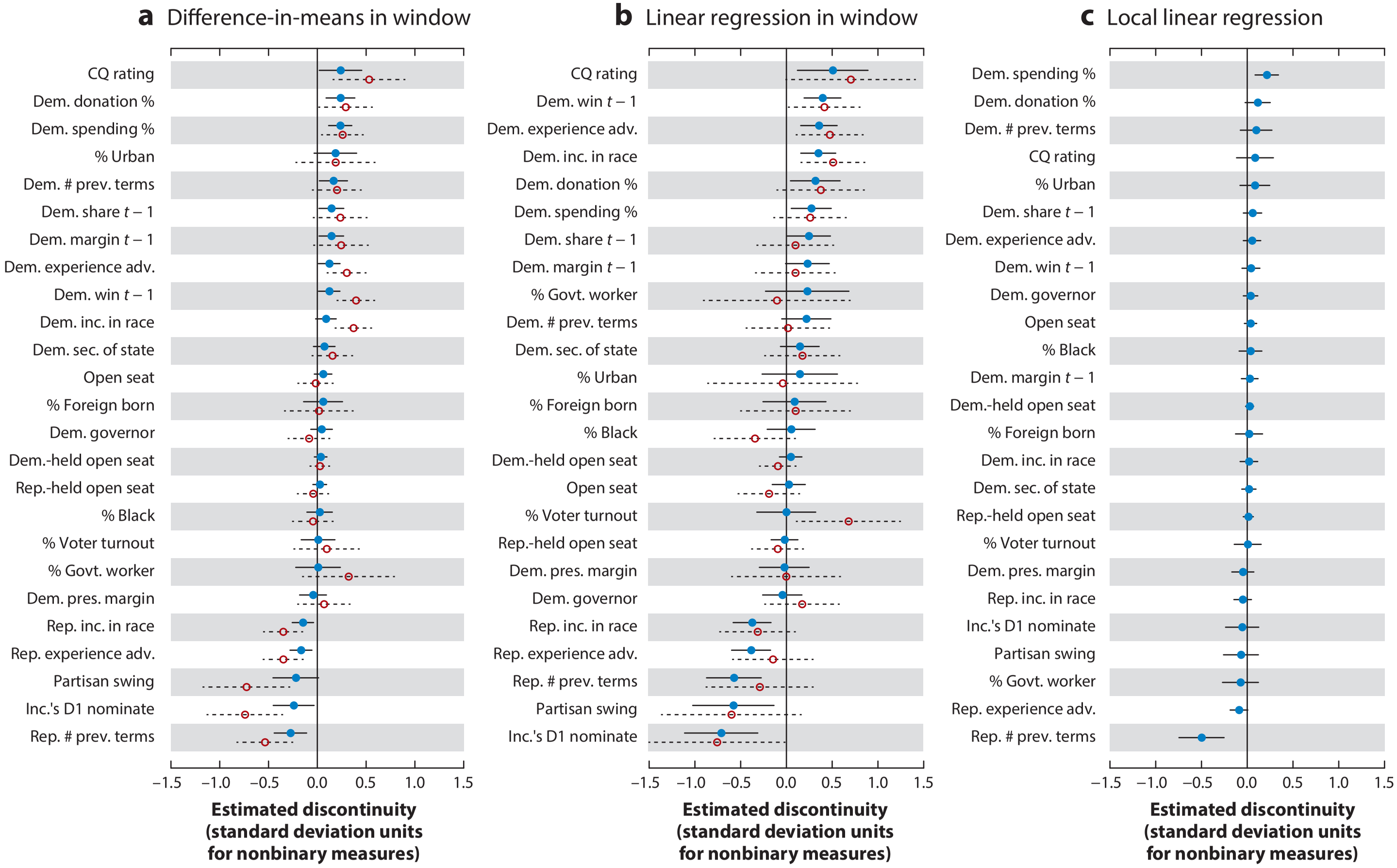

Figure 2

Comparison of estimated discontinuities in pretreatment covariates across three methods. Solid and dashed lines in each panel represent 95% confidence intervals, not corrected for multiplicity. (*a*) Filled blue circles represent estimates based on the difference-in-means estimator within the 2-percentage-point window on either side of the threshold; open red circles represent estimates within the one-half-percentage-point window. Panel (*b*) shows the estimates based on the linear regression in the same sets of windows. Panel (*c*) presents the estimates based on the local linear regression proposed by Calonico et al. (2014). Abbreviations: adv., advantage; CQ, *Congressional Quarterly*; Dem., Democratic; govt., government; inc., incumbent; pres., president; prev., previous; Rep., Republican; sec., secretary; t, time period.

controlling the false discovery rate. When moving to the one-half-percentage-point window for the differences estimator, significant discontinuities are estimated for 10 variables after multiple-testing corrections, twice as many as in the 2-percentage-point window. For the linear regression (**Figure 2*b***) in the 2-percentage-point window, 11 of 12 statistically significant estimates survive the multiple-testing correction. These results appear to strongly indicate that candidates who win close elections garner a greater proportion of overall donations, spend more money, and have more previous experience in office.

Unfortunately, the strength of the statistical evidence for sorting depends almost entirely on the method with which the discontinuity is estimated. Within the one-half-percentage-point window, the linear regression estimates only five statistically significant discontinuities, and none survive the multiple-testing correction. For the method based on the local linear regression, only two significant discontinuities are estimated, both of which survive multiple testing. The fact that the local linear regression, our most recommended method, finds much weaker evidence for sorting cannot be attributed to a lack of statistical power. In fact, across the methods examined here, the local linear regression appears to be the most powerful estimator, generally yielding shorter confidence intervals.

The results presented here illustrate how the methodological choice of different estimators and different estimation windows can affect the strength of empirical evidence. From this analysis alone, it is difficult to determine whether sorting exists in the US House elections. As discussed above (When Is the Continuity Assumption Violated?), it is unrealistic to expect campaigns to forecast the vote share with extreme precision and deploy exactly the resources necessary to win the election. It may be that the variables for which statistically significant discontinuities are estimated are correlated with postelection sorting behavior. Another possibility is a large number of missing values that exist in these data. For example, more than 50% of the observations are missing the measures for the Democratic share of total campaign spending and donations. If the missing data mechanism is related to election outcomes, this may result in false discoveries of discontinuities.[12]

Finally, we emphasize that any multiple-testing correction procedure is not a substitute for substantive judgments. For example, adding irrelevant, and hence noisy, covariates can increase the number of hypotheses tested, thus reducing the probability of rejecting the null hypotheses. For this reason, it is critical to focus on substantively important covariates. We also must remind ourselves that the failure to reject a null hypothesis does not necessarily prove its validity. Instead, it may be that we do not have enough statistical power to reject null hypotheses. Despite these caveats, our empirical analysis suggests that under the continuity assumption the empirical evidence for the lack of validity of the RD design is considerably weakened when compared to testing under the local randomization assumption.

EXTERNAL VALIDITY

The strong internal validity of the RD design comes with poor external validity. As explained above (The Continuity Assumption Does Not Imply the Local Randomization Assumption), the external validity of the RD design is limited because the average treatment effect is identified only at the threshold. In the application to close elections, this means that the incumbent party advantage can be estimated only if the election outcome is an exact tie. Because the probability

[12] In estimating the discontinuities, we follow Caughey & Sekhon (2011) in using a pairwise complete-data approach by removing only observations that have missing values on the covariate of interest and the Democratic margin. In practice, this means that different elections and districts are used to estimate the discontinuities across covariates. In addition, a complete-data analysis through list-wise deletion is not a desirable approach because it yields a total sample of only 68 observations.

of exact ties is negligibly small in most electoral settings of interest,[13] the RD design is of limited use unless incumbency advantage is constant across districts.[14] This implies that at the minimum, researchers need to extrapolate their RD estimates beyond the discontinuity threshold to relatively close elections.

To overcome this limitation, researchers, whether explicitly or implicitly, extrapolate their RD estimates away from the threshold to consider elections where a candidate prevails by a larger margin. As briefly discussed above, however, such extrapolation is credible only to the extent that we believe the relationship between the *observed* outcome and the forcing variable on one side of the threshold continues to hold for the relationship between the *counterfactual* outcome and the forcing variable on the opposite side of the threshold. Because there are no observed data on the other side of the threshold other than the outcome under the opposite treatment status, the inference hinges on an untestable assumption. This is true regardless of the approach used. The farther away from the threshold researchers extrapolate, the less credible the stable relationship assumption will be.

We illustrate the extrapolation problem by analyzing the data of Hainmueller et al. (2015), who estimate the incumbent party effect for elections to statewide office, such as governor and secretary of state, from 1946 to 2012. For comparison, we also conducted a similar analysis on the House dataset of Caughey & Sekhon (2011). The authors regressed the Democratic margin at time $t + 1$ (the outcome variable Y) on the Democratic margin at time t (the forcing variable X) and a set of additional pretreatment covariates Z, including the Democratic margin at time $t - 1$, an indicator for midterm elections, and normal vote share at time $t - 1$ and $t - 2$.[15] They estimated the regression separately on each side of the window threshold c within the window $[c_0, c_1]$. Thus, the estimators are similar to the ones given in Equations 8 and 9 except that the additional covariates Z are included. Formally, the estimators are given by

$$(\hat{\alpha}_0, \hat{\beta}_0, \hat{\gamma}_0) = \underset{\alpha_0, \beta_0, \gamma_0}{\arg\min} \sum_{i=1}^{n} 1\{c_0 \le X_i \le c\}\{Y_i - \alpha_0 - \beta_0(X_i - c) - \gamma_0^\top Z_i\}^2. \qquad 15.$$

and

$$(\hat{\alpha}_1, \hat{\beta}_1, \hat{\gamma}_1) = \underset{\alpha_1, \beta_1, \gamma_1}{\arg\min} \sum_{i=1}^{n} 1\{c < X_i \le c_1\}\{Y_i - \alpha_1 - \beta_1(X_i - c) - \gamma_1^\top Z_i\}^2. \qquad 16.$$

To visualize the conditional association between Y and X given Z, **Figure 3** presents partial residual plots for House elections and the elections for statewide offices. We first regress Y on X and Z on the each side of the threshold within the window $[-0.15, 0.15]$ as in the original analysis. Then, we plot the partial residuals $Y - \gamma_t^\top Z$ (vertical axis) based on each regression against X (horizontal

[13] Exact ties do happen. In 2015, Blaine Eaton II won a Mississippi House District 79 seat by lottery after he and his opponent received exactly 4,589 votes. However, he was later removed from office by a Republican-controlled House on the grounds that several votes in his favor should not have been counted. In the end, therefore, the winner of this election was not determined by lottery.

[14] Although Lee & Lemieux (2010) show that the average treatment effect at the threshold can also be interpreted as a weighted average treatment effect across various subpopulations, the weights are not identifiable from the observed data without additional assumptions. Therefore, in practice, it remains difficult to generalize the RD estimates to a meaningful population.

[15] Hainmueller et al. (2015) consider several sets of covariates, including the $t - 2$ lag Democratic margin variable. We omit this variable from our analyses to increase the total number of observations. Results are substantively similar even if we include this variable.

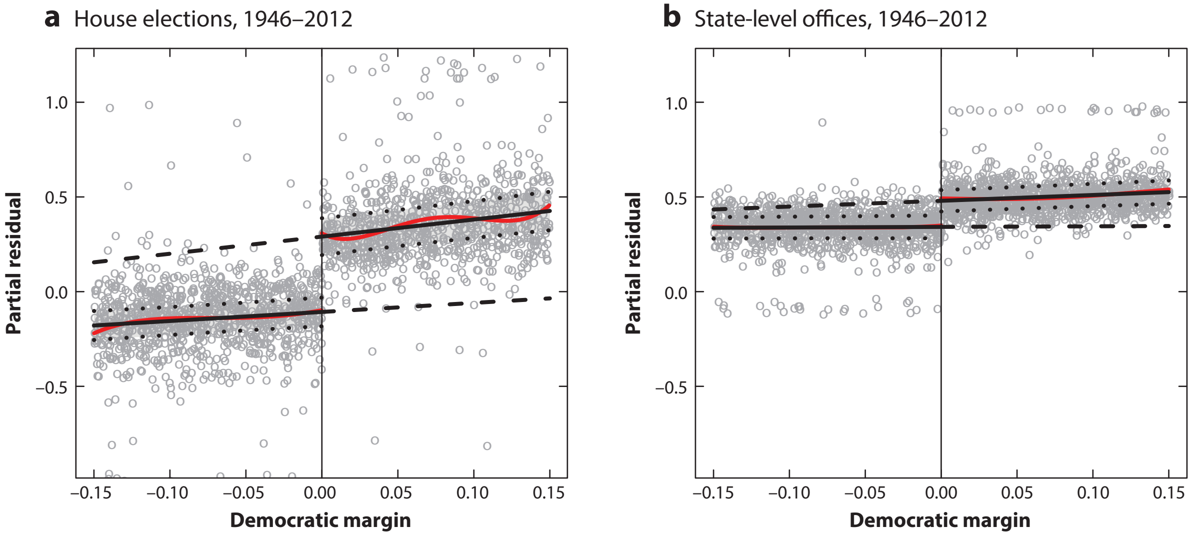

Figure 3

Extrapolation under the regression discontinuity design. The figure presents partial residual plots for (*a*) elections for the US House of Representatives and (*b*) elections for statewide offices. The solid lines correspond to the estimated conditional expectation function using observations to the left of the threshold and those to the right. The dashed black lines represent the extrapolation in the regions where no observations exist. The solid black lines are based on the linear regression fitted in the [−0.15, 0.15] window, whereas red lines represent the local linear regression estimates using all the data. The dotted black lines represent the 95% confidence intervals for the fitted values of the linear regression. The outlier partial residuals in panel (*b*) occur as a result of races that flip from competitive to uncontested. Their low frequency and unusual nature is such that they are poorly predicted by the model.

axis). The solid black lines correspond to $\hat{\alpha}_t + \hat{\beta}_t X$ from the original regression, whereas the dotted black lines represent the 95% confidence intervals based on the original regressions.

The dashed lines show the extrapolation, and the difference between the dashed and solid lines equals the estimated average treatment effect at a given value of the Democratic margin X.[16] If the slopes β_t are different, the estimated average treatment effects will change as a function of the forcing variable X. There are no observed data that can be used to empirically determine an appropriate functional form in the area of extrapolation, since all units to the left are barely-losers and all to the right barely-winners. Indeed, it is precisely this feature—the lack of overlap between treated and control groups in the forcing variable as a result of completely deterministic treatment assignment—that gives the sharp RD its name.[17] As a result, researchers must assume that the forcing variable has the same relationship with the counterfactual outcome as with the observed outcome. To make this point explicit in **Figure 3**, we also add the local linear regression fit to the partial residuals as solid black lines on each side of the threshold. As reflected by the size of the confidence intervals and, for the right side of the House data, nonlinearity in the local linear regression, there is some uncertainty about the functional form of the regression model even in

[16] For the US House elections, the initial sample (prior to removing observations with missing covariates) is identical to that used by Caughey & Sekhon (2011). For statewide offices, we use the sample analyzed by Hainmueller et al. (2015).

[17] We make the distinction between sharp and fuzzy RD designs here because, in the latter, there will be some untreated units with values of the forcing variable equal to or greater than the values for some treated units, and there will be some treated units with values of the forcing variable equal to or less than the values for some untreated units.

the area where observed data exist. The extrapolation, therefore, must be done in the presence of this uncertainty.

Angrist & Rokkanen (2015) propose a way to move forward by invoking a so-called conditional independence assumption. Within a researcher-specified window, the potential outcomes are assumed to be independent of the forcing variable X, conditional on a set of pretreatment covariates Z. Typically, this assumption is made within a window of certain size around the threshold $[c_0, c_1]$. The conditional independence assumption is

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp X_i | Z_i, c_0 \leq X_i \leq c_1. \qquad 17.$$

In terms of **Figure 3**, the assumption implies that the regression lines are flat on both sides of the threshold within the window.

The conditional independence assumption is similar to the local randomization assumption given in Equation 4 except that it is made conditional on a set of pretreatment covariates Z. In addition, whereas the local randomization assumption implies that the treatment assignment is randomized, the conditional independence assumption states that the forcing variable is independent of the potential outcomes. Despite these differences, the conditional independence assumption implies that the average potential outcome is independent of the forcing variable conditional on Z. This is similar to the result obtained under the local randomization assumption except that we now condition on Z_i (see Equation 5).

To apply this method, researchers must decide how far away from the threshold the RD estimates should be extrapolated by determining the window size. Angrist & Rokkanen (2015) suggest a decision rule based on whether or not the estimated slope coefficient for the forcing variable $\hat{\beta}_t$ is statistically significantly different from zero. The authors propose starting in a narrow window around the cutpoint and increasing the window size incrementally until the p-value of the coefficient on the forcing variable is greater than a predetermined level. That is, they propose finding the largest window in which the forcing variable, conditional on the control covariates Z, is not statistically significantly associated with the outcome. Applying this approach to close elections, Hainmueller et al. (2015) find that the RD estimates can be extrapolated as far as 15 percentage points from the threshold in elections for statewide offices.

Figure 4 presents the results for elections for the US House and statewide offices. Each plot displays the estimated coefficient for the forcing variable $\hat{\beta}_t$ and its 95% confidence intervals for various window sizes (horizontal axis). The slope coefficient is estimated to be substantively large, though not necessarily statistically significant, near the threshold for the US House elections. This is consistent with the finding (discussed in the previous section) that the association between the outcome and forcing variables is strong near the threshold in the US House elections. For the statewide offices, the association is weaker, although the estimated coefficient appears to increase near the threshold. In general, one must be careful about the use of p-values to determine the window size. For example, even though the estimated coefficient very near the threshold is not statistically significantly different from zero in both cases, the estimated coefficient is substantively large, and a large p-value may be simply due to a lack of statistical power rather than a lack of association.

More importantly, there is no methodological reason to prefer the conditional independence assumption over the continuity assumption, which can also be made conditional on the set of additional covariates.[18] In fact, from a purely statistical perspective, the conditional independence

[18] We should note that the continuity assumption conditional on covariates is actually stronger than the continuity assumption without such conditioning. We thank Rocío Titiunik for alerting us to this subtle but important point.

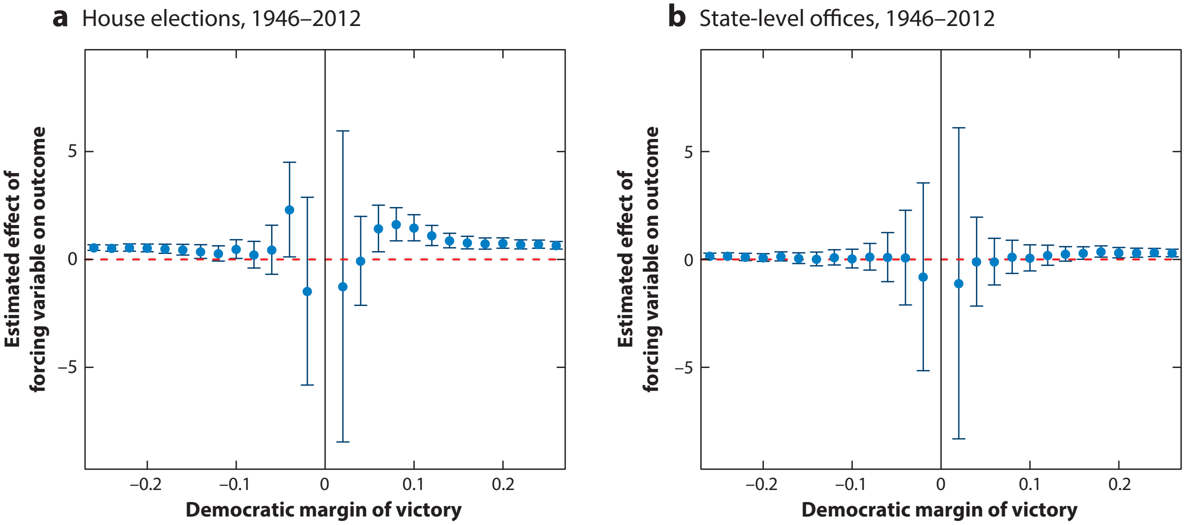

Figure 4

Validity of the conditional independence assumption in the elections for the US House of Representatives and statewide offices, 1946–2008/2012. Plots present, using various window sizes, the estimated coefficient (along with 95% confidence intervals) of the forcing variable, Democratic margin of victory in the election at time t, in the regression model whose outcome variable is the Democratic margin in the election at time $t + 1$. Panel (*a*) presents the results for US House elections. Panel (*b*) presents the same analysis applied to elections for the statewide offices. The model for both analyses also includes a set of pretreatment covariates as additional predictors. It appears that for House elections there is a strong association, though not always statistically significant, between the outcome and forcing variable even in a narrow window near the threshold.

assumption is more restrictive because it unnecessarily requires the slope coefficient γ_t for the forcing variable to be zero conditional on the pretreatment covariates. It may be preferable to estimate the slope coefficient based on the observed data and then extrapolate using the fitted lines with nonzero slopes as in **Figure 3**. Because no data can inform how extrapolation should be conducted, the approach must be justified on substantive grounds. At this point, the internal validity of the RD design essentially reduces to that of standard observational studies: Researchers must rely on a substantive (but untestable) argument that they are able to measure a sufficient set of pretreatment covariates Z, and hence the Democratic margin of the current election is no longer predictive of the outcome of the next election.

In summary, the lack of external validity is an important limitation of the RD design. Generalization of RD estimates necessitates extrapolation, which in turn rests on an untestable assumption similar to the one made in standard observational studies. This is important because in the case of close elections, the causal estimand under the RD design may not be of interest unless it is applicable to elections beyond those at the discontinuity threshold. Because no data exist to support extrapolation away from the cutpoint, however, researchers should, at the minimum, provide a substantive argument for why the conditional expectation function of observed outcome for the treated (or untreated) units is likely to be a valid approximation of the unobserved relationship between the forcing variable and the counterfactual outcome for those units. Clearly, a better approach is to use data when justifying the assumption underlying the extrapolation. Good examples of such efforts include Wing & Cook's (2013) proposal to use pretest data and Cattaneo et al.'s (2015b) exploitation of the existence of multiple discontinuity thresholds.

BEYOND CLOSE ELECTIONS

In this article, we clarified several misunderstandings about the RD design in the study of close elections by revisiting the controversy in the literature. Under the RD design, the strength of internal validity originates from the fact that the usual as-if-random assumption in observational studies is not necessary. Instead, only the continuity assumption is required for the identification of average causal effects at the discontinuity threshold. In the context of close elections, this suggests that postelection manipulation, including election fraud, rather than incumbents' structural advantages in campaign resources, are needed for the violation of this assumption. We have also shown that the difference between the local randomization and continuity assumptions can affect the choice of estimation methods and therefore one's empirical conclusions.

A major limitation of the RD design, however, is that the average causal effect is identified only at the threshold. This quantity is rarely of interest because researchers are unlikely to learn much by estimating incumbency advantage for districts with exact tie votes. To address this limitation, they must extrapolate the RD estimates away from the threshold. We have shown that such extrapolation must be justified by a strong substantive argument. The reason is that no observed data can inform the validity of extrapolation and an untestable assumption must be invoked. Therefore, the RD design sharply illustrates the trade-off between internal and external validity. When only the minimal identification assumption is made, the RD design has strong internal validity but almost no generalizability. Once the RD estimates are extrapolated away from the threshold, any resulting estimates lose the strong internal validity for which the RD design is known.

These methodological issues are also relevant for other applications of the RD design. Consider the temporal RD design where plausibly exogenous events such as a terrorist attack (Legewie 2013) or an election result (Pierce et al. 2016) expose units to a treatment at a specific time point. As a result, all units after this time point belong to the treatment group, and none before it does. As in the close elections case, the average causal effect at the moment of the event is identified, but this quantity is of limited substantive value. Most researchers, therefore, extrapolate RD estimates into future time periods where the counterfactual outcome can never be observed. Such inference relies on the (untestable) assumption that the conditional expectation function for the control group can be used to approximate the counterfactual for the treated group, significantly sacrificing its internal validity.

As the applications of the RD design become increasingly popular, other methodological issues that are not discussed in this article also arise. Consider the case of geographic RD, where administrative and other boundaries serve as the thresholds (see e.g., Keele & Titiunik 2015, 2016). This RD design has a two-dimensional threshold, which poses an additional modeling challenge. Moreover, the substantive interpretation of RD estimates is complicated by the fact that at administrative boundaries many factors are likely to change. For example, in his study of the influence of relative group size on ethnic relations, Posner (2004) uses the Zambia–Malawi border as a geographic discontinuity, creating different relative group sizes on either side of the border. Yet, the border itself is the source of many treatments, including different colonial and postcolonial histories. This bundle of several treatments means that only through a careful substantive discussion of these legacies can we make a case for the importance of differing relative group size.[19] Finally, because scholars are unlikely to be interested in the causal effects at the administrative boundaries, extrapolation must also be done in this spatial setting to generalize the RD estimates.

Despite these challenges, we believe that the RD design is a potentially useful tool for estimating causal effects in observational studies. Recent methodological developments have also made it possible for researchers to draw valid causal inferences in a principled and transparent manner.

[19] In fact, a substantial portion of Posner (2004) is devoted to precisely this exercise.

Nevertheless, as demonstrated in this article, the proper application of the RD design demands the correct understanding and interpretation of its identification assumption as well as the careful choice and implementation of estimation methods. In addition, when inference depends on an untestable assumption, as required by the generalization of RD estimates beyond the threshold, researchers must, at the minimum, present a strong substantive defense of the assumption.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

Comments from Naoki Egami, Andy Eggers, Julia Morse, Rocío Titiunik, and Teppei Yamamoto have improved the manuscript.

LITERATURE CITED

Angrist J, Rokkanen M. 2015. Wanna get away? RD identification away from the cutoff. *J. Am. Stat. Assoc.* 110(512):1331–44

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* 57(1):289–300

Benjamini Y, Yekutieli D. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 29(4):1165–88

Boas T, Hidalgo D, Richardson N. 2014. The spoils of victory: campaign donations and government contracts in Brazil. *J. Polit.* 76(2):415–29

Butler DM. 2009. A regression discontinuity design analysis of the incumbency advantage and tenure in the U.S. House. *Elect. Stud.* 28(1):123–28

Calonico S, Cattaneo M, Titiunik R. 2014. Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* 82(6):2295–326

Calonico S, Cattaneo M, Titiunik R. 2015. `rdrobust`: an `R` package for robust nonparametric inference in regression-discontinuity designs. *The* `R` *Journal* 7(1):38–51

Cattaneo MD, Frandsen B, Titiunik R. 2015a. Randomization inference in the regression discontinuity design: an application to party advantages in the U.S. Senate. *J. Causal Inference* 3(1):1–24

Cattaneo MD, Keele L, Titiunik R, Vazquez-Bare G. 2015b. *Interpreting regression discontinuity designs with multiple cutoffs*. Work. pap., Univ. Mich., Penn. State Univ.

Cattaneo MD, Titiunik R, Vazquez-Bare G. 2015c. *Comparing inference approaches in RD designs: a reexamination of the effect of Head Start on child mortality*. Work. pap., Univ. Mich.

Caughey D, Sekhon JS. 2011. Elections and the regression discontinuity design: lessons from close U.S. House races, 1942–2008. *Polit. Anal.* 19(4):385–408

Dinas E. 2014. Does choice bring loyalty? Electoral participation and the development of party identification. *Am. J. Polit. Sci.* 58(2):449–65

Dunning T. 2008. Improving causal inference: strengths and limitations of natural experiments. *Polit. Res. Q.* 61(2):282–93

Dunning T, Nilekani J. 2013. Ethnic quotas and political mobilization: caste, parties, and distribution in Indian village councils. *Am. Polit. Sci. Rev.* 107:35–56

Eggers A, Fowler A, Hainmueller J, Hall A, Snyder J. 2015a. On the validity of the regression discontinuity design for estimating electoral effects: new evidence from over 40,000 close races. *Am. J. Polit. Sci.* 59(1):259–74

Eggers AC, Freier R, Grembi V, Nannicini T. 2015b. *Regression discontinuity designs based on population thresholds: pitfalls and solutions*. DIW Berlin Disc. Pap. No. 1503

Eggers AC, Hainmueller J. 2009. MPs for sale? Returns to office in postwar British politics. *Am. Polit. Sci. Rev.* 103(4):513

Erikson R, Folke O, Snyder J. 2015. A gubernatorial helping hand? How governors affect presidential elections. *J. Polit.* 77(2):491–504

Erikson R, Rader K. 2013. *Much ado about nothing: RDD and the incumbency advantage*. Presented at Annu. Eur. Polit. Sci. Assoc. Conv., 3rd, Barcelona, June 20–22

Fan J, Gijbels I. 1996. *Local Polynomial Modelling and Its Applications*. Boca Raton, FL: Chapman & Hall/CRC

Folke O, Snyder J. 2012. Gubernatorial midterm slumps. *Am. J. Polit. Sci.* 56(4):931–48

Fouirnaies A, Hall A. 2014. The size and sources of the financial incumbency advantage: evidence from American legislatures. *J. Polit.* 76(3):711–24

Friedman J, Holden R. 2009. The rising incumbent reelection rate: What's gerrymandering got to do with it? *J. Polit.* 71(2):593–611

Galasso V, Nannicini T. 2011. Competing on good politicians. *Am. Polit. Sci. Rev.* 105(1):79–99

Gerber AS, Kessler DP, Meredith M. 2011. The persuasive effects of direct mail: a regression discontinuity based approach. *J. Polit.* 73(01):140–55

Gerber ER, Hopkins DJ. 2011. When mayors matter: estimating the impact of mayoral partisanship on city policy. *Am. J. Polit. Sci.* 55(2):326–39

Grimmer J, Hersh E, Feinstein B, Carpenter D. 2011. *Are close elections random?* Work. pap., Stanford Univ., Yale Univ., Harvard Univ.

Hainmueller J, Hall AB, Snyder JM. 2015. Assessing the external validity of election RD estimates: an investigation of the incumbency advantage. *J. Polit.* 77(3):707–20

Holbein J, Hillygus S. 2016. Making young voters: the impact of preregistration on youth turnout. *Am. J. Polit. Sci.* In press

Imbens G, Kalyanaraman K. 2012. Optimal bandwidth choice for the regression discontinuity estimator. *Rev. Econ. Stud.* 79(3):933–59

Imbens GW, Lemieux T. 2008. Regression discontinuity designs: a guide to practice. *J. Econom.* 142(2):615–35

Keele L, Titiunik R. 2015. Geographic boundaries as regression discontinuities. *Polit. Anal.* 23(1):127–55

Keele L, Titiunik R. 2016. Natural experiments based on geography. *Polit. Sci. Res. Methods* 4(1):65–95

Krasno JS, Green DP. 2008. Do televised presidential ads increase voter turnout? Evidence from a natural experiment. *J. Polit.* 70(1):245–61

Lee D, Lemieux T. 2010. Regression discontinuity designs in economics. *J. Econ. Lit.* 48:281–355

Lee D, Moretti E, Butler M. 2004. Do voters affect or elect policies? Evidence from the US House. *Q. J. Econ.* 119(3):807–59

Lee DS. 2008. Randomized experiments from non-random selection in U.S. House elections. *J. Econom.* 142(2):675–97

Legewie J. 2013. Terrorist events and attitudes toward immigrants: a natural experiment. *Am. J. Sociol.* 118(5):1199–245

McCrary J. 2008. Manipulation of the running variable in the regression discontinuity design: a density test. *J. Econom.* 142(2):698–714

Pierce L, Rogers T, Snyder JA. 2016. Losing hurts: the happiness impact of partisan electoral loss. *J. Exp. Polit. Sci.* In press

Posner DN. 2004. The political salience of cultural difference: why Chewas and Tumbukas are allies in Zambia and adversaries in Malawi. *Am. Polit. Sci. Rev.* 98(4):529–45

Samii C. 2013. Perils or promise of ethnic integration? Evidence from a hard case in Burundi. *Am. Polit. Sci. Rev.* 107(3):558–73

Skovron C, Titiunik R. 2015. *A practical guide to regression discontinuity*. Work. pap., Dep. Polit. Sci., Univ. Mich.

Snyder J. 2005. *Detecting manipulation in U.S. House elections.* Work. pap. 2–15, Haas School Bus., Univ. Calif. Berkeley

Thistlewaite D, Campbell D. 1960. Regression-discontinuity analysis: an alternative to the ex-post-facto experiment. *J. Educ. Psychol.* 51(6):309–17

Wing C, Cook TD. 2013. Strengthening the regression discontinuity design using additional design elements: a within-study comparison. *J. Policy Anal. Manag.* 32(4):853–77

Annual Review of Political Science

Volume 19, 2016

Contents

Indexes

Errata

An online log of corrections to *Annual Review of Political Science* articles may be found at http://www.annualreviews.org/errata/polisci