Preregistration for "Agency Versus Structure in International Affairs: Evidence from Electoral Discontinuities"

Andrew Bertoli, Allan Dafoe, and Robert Trager

5 May 2016

Abstract

This document summarizes the research objectives for our study on the effect of leaders on foreign policy. The main goal of this project is to shed light on the longstanding question of whether leaders have an important, independent impact on international relations. We will investigate this question using regression discontinuity. Specifically, we will look at cases where competing presidential candidates with different traits narrowly won and lost close elections. In the following pages, we discuss the research question and existing literature in more detail, describe our data and research design, and summarize the primary variables and statistical tests that we will use in our paper. The purpose is to discipline our data analysis and reduce the risk of multiple comparisons bias.

1 Introduction

To what extent do leaders shape foreign policy? Do they have a large independent effect on state behavior, or is their influence largely constrained by structural factors. This question matters for both theory and policy. On the theoretical side, knowing the degree to which leaders shape important international developments is key to recognizing the contingency of history and the weight that we should place on individuals when analyzing the past. On the policy side, the U.S. government often tries to remove leaders who undermine international cooperation, through peaceful and occasionally violent means. Understanding the independent effect of leaders is important for anticipating foreign policy behavior of countries and assessing the extent to which removing leaders would promote a state's interests.

Despite the importance of this question, our understanding remains highly uncertain. There are prominent debates in IR and history over whether individual leaders have a meaningful impact on foreign policy (Byman and Pollack 2001). A major challenge facing this research program is that the kinds of people who become leaders is heavily shaped by other factors, and leader's attitudes are similarly strongly influenced by other factors. Therefore, the leaders who seem particularly hostile or conciliatory may just be reflecting the broader preferences of their countries, and therefore they may not be exerting much of an independent effect on international relations. Unfortunately, there is no way to conduct an experiment where leaders are randomly assigned to countries.

While doing a real experiment is impossible, we plan to investigate a series of plausible natural experiments that will shed light on the importance of leaders for foreign policy. These natural experiments will involve looking at cases where certain presidential candidates barely defeated their opponents in close democratic elections from 1815-2010. Given the randomness in large national elections, it is often plausibly as-if random which candidate became president in cases where both candidates were on the verge of winning. Therefore,

this design will allow us to estimate the effect of electing candidates with particular traits on foreign policy.

The purpose of preregistration is to articulate, to the extent feasible, the traits that we intend to examine and the statistical tests that we will use. By doing so, we discipline the data analysis, reducing the risk of multiple comparisons bias. After all, the number of attributes that we could potentially investigate is unlimited. We chose our traits by first identifying what leader attributes other researchers have hypothesized matter in past research. We then narrowed that list down to characteristics that we could feasibly collect data on. After collecting the data, we then found the attributes that seemed testable based on the set of close elections that we had put together. For example, there were not enough close elections involving women for us to use gender as a trait, although we were interested in addressing that question. For our statistical methods, we chose our tests by conducting power analysis, which are summarized in Section 6.

Preregistration should not be regarded as a binding commitment. Doing so would either excessively constrain researchers from analyzing the data in the most appropriate way, or would unnecessarily deter them from preregistering. The perfect should not be the enemy of the good. Instead, preregistration should be understood as an expression of one's motivation and analysis plan going into a study, as detailed as is optimal given the trade-offs.

If new information or new understanding leads us to prefer another analysis strategy, we will follow that, explaining our reasoning. To the extent that we deviate, we intend to decide specification on principle and blind to the outcome. Comparisons not anticipated in this document should be regarded as exploratory, whereas those precisely anticipated can be regarded as confirmatory; many comparisons will be of an intermediate "gray" type.[1] We may also update this preregistration as we decide on new analysis plans, consistent with the "track changes" model of preregistration.[2]

Replication files (with data and complete code) will be shared so that others can assess whether any patterns that we identify are robust to other reasonable specifications (ideally those replication analyses should also be preregistered).

2 Existing Literature

Expectations about the Effect of Age on Conflict

Past predictions about the effect of age are varied. On the one hand intuition and a plausible physiological theory suggest youth would be associated with more aggressive policies. Many international conflicts are said to derive from the desire of youth to overturn the current

[1]For discussion of this and "standard operating procedures" for conditions outside a pre-analysis plan, see Green, D., & Lin, W. (n.d.). Standard Operating Procedures: A Safety Net for Pre-Analysis Plans. PS: Political Science and Politics. and Humphreys, Macartan, Raul Sanchez de la Sierra, and Peter van der Windt. 2013 "Fishing, Commitment, and Communication: A Proposal for Comprehensive Nonbinding Research Registration." Political Analysis 21 (1): 1-20.

[2]Bidwell, Kelly, Katherine Casey, and Rachel Glennerster. 2015a. "The Impact of Voter Knowledge Initiatives in Sierra Leone." AEA RCT Registry. https://www.socialscienceregistry.org/trials/26.

order, from Alexander the Great who was only 33 when he died to the 23 year old Franz Josef rejecting the advice of the 80 year old Metternich to break with Russia during the Crimean War. Physiologically, many studies associate testosterone with aggression (e.g. Archer 1991, 2006, Scerbo 1994), and testosterone is known to decline predictable in men after the mid 20s (Schatzl et al. 2003, Seidman 2003, Juul and Skakkebaek 2002).

On the other hand, however, recently analyses of the effect of age in international politics have found that age is in fact associated with a dramatically increasing likelihood of international conflict. Horowitz, McDermott and Stam (2005) estimate that a 20 year change from 40 to 60 doubles the risk of initiating a militarized interstate dispute, with the effect of age on the likelihood of conflict increasing at an even faster rate thereafter. Horowitz, Stam and Ellis (2015) estimate an approximate 18% increase in the probability of conflict for each 10 year increase in age (from a baseline prediction of 10% likelihood of conflict initiation in a given dyad-year). Interestingly, they also find that these effects are stronger for democracies than autocracies even though the argument, and finding, in Horowitz, McDermott and Stam (2005) is the opposite: age is predicted to have a larger effect in an autocracy because of the relative freedom of the leader. In both works, the authors argue that the effect of age on conflict stems from the enhanced ability of older leaders to make war due to greater power consolidation and their shorter time horizons of older leaders. One might equally argue, however, that a shorter time horizon would make a leader less willing to bear the short-term costs of conflict for the long-term benefits of victory. Further, as Horowitz, McDermott and Stam (2005, 681) note: "older leaders may have longer time horizons than younger leaders because they are more likely to have children."

Because studies in international relations analyze observational data, they are susceptible to selection biases. In fact, examination of specific cases suggests strongly that such selection is present. Winston Churchill was brought as Prime Minister of Great Britain shortly after the start of the Second World War. Then, despite overwhelming popularity as a result of his successful prosecution of the conflict, he was voted out of office shortly after victory was achieved by the Allies in Europe. The British people judged him the best leader to fight the war, but not the best candidate to lead Britain in peacetime. In fact, he was succeeded by the decade younger Clement Attlee.

Thus, a plausible alternative explanation for the observed relationship between age and conflict, particularly in democracies, is that older leaders are selected into office in dangerous times. If this is true, then these observational studies fail to identify the causal effect of age. We therefore conclude that the electoral regression discontinuity design may identify a very different effect of age, more consistent both with intuition and with plausible theory derived from physiology. In particular, we intend to examine the impact of age on a) MID initiation and b) change in MID frequency.

2.1 Expectations about the Effect of Incumbency on Foreign Policy

Substantial recent scholarship has focused on whether different types of leaders behave differently in similar international political contexts (**?**, **?**, Hermann et al. 2001). Would the constraints of the system cause President Trump to make the same decisions in office as President Clinton? If not, how do leaders' experiences and genetics interact with political context to influence decision-making? Anecdotal evidence suggests that the influence of individuals,

distinct from the system structure, is substantial, but a long tradition in international relations scholarship focuses on the international system as the most important determinant of international outcomes (**?**, **?**). Domestic political dynamics provide incentives for individuals with divergent preferences to behave similarly as Presidents and Prime Ministers. Leaders with hawkish reputations have an easier time "offering the olive branch," while leaders that are perceived as more dovish sometimes find it politically easier to take their countries into a war. Politicians have incentives to play to type in legislatures, and to act against type when they hold executive office (**?**). A result may be policy consistency of executives across leaders and political parties.

While some literatures have demonstrated correlations between leader attributes and international outcomes, the possibility remains that international context selects leaders with certain qualities. Horowitz, Stam and Ellis (2015) show, for instance, that the most violent moments in the international system can be accurately predicted through both systemic variable and leader attribute models. Since many system variables in these models are not influenced by the particular leaders in office, a primary candidate for the explanation for the equal effectiveness of the two classes of models must be that the international system selects leaders of particular types in particular contexts.

The RD design addresses this issue, allowing us to identify the independent causal effect of a change in leaders on a change in foreign policies. In particular, we will examine the effect on a) an absolute change in MID frequency, b) an absolute change in alignment relationships (evaluated through s-scores).

2.2 Expectations about the Effect of Ideology on Foreign Policy

While there are many dimensions of difference between leaders, in many contexts, a single left-right dimension has been shown to capture most of the variance in political opinion (McCarty). The left-right dimension has also been shown to correlate with moral values (**?**). Those on the left place greater value on fairness and duties not to harm, while those on the right value the preservation of authority, loyalty to an ingroup, and the purity of sanctified objects (including religious and cultural groupings). **?** show that the values associated with left and right predict foreign policy attitudes in the United States. In particular, the conservative values predict "militant internationalism" while the liberal values predict a more cooperative approach in international affairs.[3]

We therefore expect more socially conservative leaders to be associated with more aggressive foreign policies. In particular, we will examine the effect of location on the left-right dimension on a) MID initiation and b) change in MID frequency.

Consistent with our theoretical expectations, we locate leaders on a 5-point left-right spectrum primarily according to their views at the time of the election on social questions associated with liberalism and conservatism. Candidates were judged further to the right when they expressed support for "traditional values," national, religious, racial or ethnic in-groups, the benefits of authority and traditional sources of authority such as a monarchy. Candidates were judged further to the left when they expressed inclusive sentiments, a duty of care for vulnerable groups, and support for democratic principles. Secondarily, we evaluate

[3]See also **?**, **?**, and **?**.

candidates as left or right on economic policy preferences. Advocacy for wealthier interests places a candidate further to the right and advocacy for the less well-off is associated with the left. In practice, these two social and economic coding dimensions are highly correlated with the primary exceptions come from communist and post-communist countries. In these cases, the primary social dimension determined the left-right coding on our scale.

3 Research Design

To answer this question, we will look at a set of potential natural experiments where certain leaders came to office instead of others in a manner that is plausibly as-if random. These involve comparing democratic countries being led by leaders (with a certain set of attributes, denoted X) who narrowly won against others without those attributes ($\not X$), or vice versa. This comparison gets us closer to an unbiased estimate of a causal effect that is policy and normatively relevant. The causal estimand is:

> *the effect of electing a leader with attribute X vs electing a leader without attribute X.*

amongst those country years where the previous election was likely to have been close between these two kinds of people.

There are two main assumptions required of this research design. The first, required for internal validity, is that the outcomes of very close elections are as-if random. In other words, it is not possible to "sort around the cutpoint", such as would be possible if one candidate could count the votes and add as many as was required to win. The second assumption, required for external validity, is that the countries with close elections are fairly representative of other countries, or at least other democracies. If these countries tend to be atypical, then the causal estimate that we might correctly find for them could be misleading when extrapolated to other countries.

Fortunately, we will be able to test both of these assumptions to a degree. The as-if randomness assumption can be tested (though not proven) by seeing whether the countries on either side of the cut-point are similar across pretreatment characteristics. The most important of these factors is prior levels of the outcome variables. The external validity assumption can be tested by assessing how similar the democracies in our sample are to other democracies. If they appear to be fairly representative, then there is reason to believe that our results in large part hold for other democracies.

To be clear, we are not claiming to identify the causal effect of attribute X (where X could be age, military experience, etc...). Rather, we are claiming to identify the causal effect of electing a leader with attribute X. This effect may be due to X, or may be merely correlated with X. Nevertheless, it is a policy relevant causal estimand because democratic publics are often confronted with voting for a leader. Similarly, foreign policy observers often watch elections in other countries with great interest because of the potential implications for foreign relations. Our design provides an estimate relevant to both these groups: the effect of one kind of leader being elected, as opposed to the other kind.

4 Data

4.1 Election Data

We constructed the election dataset as follows. We started by obtaining the results of all presidential elections where the top two candidates were within 10% of the cut-point (40%-60% range).[4] In this initial stage, we included both democracies and non-democracies. Most of this data is available in Bertoli (2016), which lists all elections in democracies where the most powerful party was close to achieving united government. However, this dataset was missing some close presidential elections where no party was close to united government or where the state was a non-democracy. The remaining data was available in the following sources: *Elections in the Americas: A Data Handbook* (2005), *Elections in Europe: A Data Handbook* (2010), *Elections in Africa: A Data Handbook* (1999), and *Elections in Asia and the Pacific: A Data Handbook* (2001). We also obtained results for very recent elections in online databases, including the International Foundation for Electoral Systems' Election Guide, the African Election Database, and the European Election Database.

We also coded for whether each country was a democracy in the election year. We counted a country as a democracy in a given year if it had POLITY IV Institutionalized Democracy Score above five, which is consistent with past research (Schrock-Jacobson 2012; Marshall, Gurr, and Jaggers 2013). We distinguished between democracies and non-democracies because it is plausible that close elections in non-democracies are non-random. While there are good theoretical reasons for believing that close democratic elections should be as good as random, there are also reasons for thinking that close elections in non-democracies might not be. Therefore, we plan to test our hypotheses primarily on democracies.

Altogether, the data consists of 245 close elections, 153 of which occurred in democracies. Not all of these cases may be used for every test, because we can only use close elections where there was variation between the candidates on the variable of interest. For instance, if we wanted to test whether there was more persistence in foreign policy variable when the incumbent party barely won the presidential election compared to when they barely lose to a challenger party, we would have to exclude all of the elections where the incumbent party did not run or was not close to winning. Therefore, 153 close elections is the upper limit of the sample size for democracies.

4.2 Leader Traits

We collected the ages of the winning and losing presidential candidate, their ideological leanings (on 5-point scale), and whether they were from the party that previously controlled government (party incumbency status). Much of the data for the winning candidates was available in the Leader Experiences, Attributes, and Decisions (LEAD) dataset (Horowitz and Stam 2016). We obtained the remaining data from online biographies of candidates. For ideological leanings, we plan to treat candidates as having the same ideology if they are within one point of each other and different if they are two or more points apart.

[4]For the U.S., we used the electoral college votes rather than the popular vote.

4.3 Dependent Variables

We gathered a number of dependent variables related to international conflict and cooperation. The first was the average number of Militarized Interstate Disputes (MIDs) that each country initiated per year over the winning candidates term in office. These disputes are instances where states explicitly threaten, display, or use force against other countries (Ghosn, Palmer, Bremer 2004), and are listed in the Correlates of War database. The second dependent variable was the number of MIDs per year that involved the use of force. This measure excludes very low-level disputes that merely involve saber-rattling.

Our choice of dependent variable largely depends on our theoretical expectations about the leader traits we are investigating. Since age is closely related to testosterone levels, we will focus on MIDs and MIDs involving force for cases where older candidates narrowly beat and lost to younger candidates.

5 Estimation Techniques

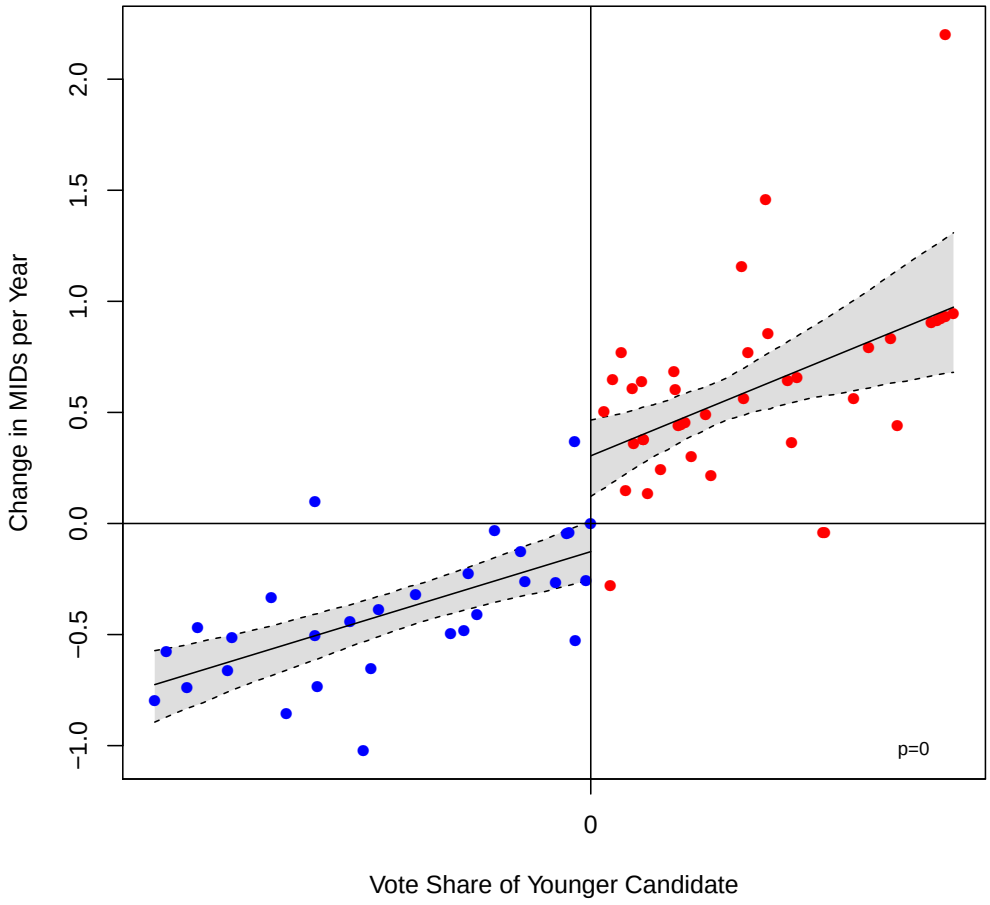
There are three standard estimation techniques that we will consider using in the paper. The first is to draw around the data close to the cut-point (48%-52%) and treat the data inside like experimental data (Dunning 2012; Cattaneo, Frandsen, and Titiunik 2015). This approach assumes as-if randomness for the units that are close to the cut-point. The statistical tests that are typically used with this approach are t-tests or permutation tests of a difference in means. We will use a t-test with permutation inference as a robustness check.

The second standard approach involves estimating a linear regression, with different slopes on either side of the cut-point and a parameter estimating the difference between the regression lines at the cut-point. Figure 1 illustrates this approach using simulated data. It does not require that election outcomes were as-if random close to the cut-point. Instead, it assumes that the potential outcomes are smooth around the cut-point and can be estimated reasonably well with a linear functional form. The advantage of this approach is that it uses all of the data, which decreases the variance of the estimator and increases power. The drawback is that it leans more on the (linear) functional form being appropriate.

The third approach is similar to the second, but it uses local linear regression. This will increase the variance of the estimator by restricting the amount of data that is used, but it also makes the analysis less dependent on modeling assumptions. To conduct this analysis, we use the "rdrobust" software developed by Calonico, Cattaneo, and Titiunik (2014). This software packages produces bias-corrected estimates with robust standard errors. The optimal bandwidth is calculated to minimize the mean-squared error, which is a common criterion for bandwidth selection.

We decide which tests we will use in our primary analysis based on how these estimation techniques perform in our power analyses using the real data. We describe and conduct these in the next section.

Figure 1 Example Linear Regression Graph with Simulated Data



6 Power Analysis

To get a better sense of how these estimation techniques would perform on our sample, we looked at how they did when applied to the real outcomes and a randomized version of the treatment. We describe these procedures for the age example here. The procedures for idealogical leanings and party incumbency will be the same.

To scramble the treatment for age, we started by creating the forcing variable, Z, which is the distance that the older candidate was to winning the election. For example, if $Z = 0.01$, the older candidate won the election by 1%, whereas if $Z = -0.05$, the older candidate lost the election by 5%. After constructing this variable, we randomly changed whether Z was positive or negative, meaning that we shuffled whether the older or younger candidate won the election. While doing this, we kept the absolute value of the forcing variable for each unit the same, so that any unit's forcing variable could only be at its original value or the opposite value. To keep the new data as much like our sample as possible, we also held fixed (1) the number of cases where Z was positive and negative in the 48%-52% range and (2) the number of cases where Z was positive and negative outside that range.

Table 1 Power Analysis: Percent of Times that the Tests Detect a Treatment Effect

All Elections With Candidates That Were At Least One Year Apart			
Treatment Effect Size	t-test (n=46)	Linear Regression (n=151)	rdrobust (n=151)
0.0 x SD	5%	4%	3%
0.3 x SD	14%	14%	4%
0.5 x SD	29%	35%	7%
0.8 x SD	65%	72%	17%
1.2 x SD	95%	98%	30%
All Elections With Candidates That Were At Least Five Years Apart			
Treatment Effect Size	t-test (n=36)	Linear Regression (n=109)	rdrobust (n=109)
0.0 x SD	5%	3%	10%
0.3 x SD	14%	7%	12%
0.5 x SD	33%	20%	17%
0.8 x SD	71%	54%	33%
1.2 x SD	98%	89%	57%
All Elections With Candidates That Were At Least Ten Years Apart			
Treatment Effect Size	t-test (n=24)	Linear Regression (n=70)	rdrobust
0.0 x SD	4%	2%	12%
0.3 x SD	10%	5%	14%
0.5 x SD	25%	18%	18%
0.8 x SD	65%	72%	17%
1.2 x SD	89%	85%	47%

For each power analysis, we then added the stipulated treatment effect for the units that surpassed the cut-point. The values of the constant treatment effect that we experimented with were {0, 0.3, 0.5, 0.8, and 1.2} times the standard deviation of the outcome variable (Change in MIDs per year), also often called Cohen's *d*. These effect sizes can be regarded as spanning small (0.3) to large (1.2) effects. We then applied the three estimation techniques to the data and checked for whether they rejected the null hypothesis of no treatment effect. To get a sense of how likely they would be to so, we re-scrambled the treatment 1,000 times and took the percent of times that the tests rejected the null.

The results from this analysis are presented in Table 1. While using all the data where the two candidates differed in age by at least one year, we found that the t-test and linear regression preformed much better than local linear regression ("rdrobust"). They both preserved correct size (had less than 5% when the effect was actually zero), and often had much greater power (greater rejection rate when the causal effect was not zero).

Moreover, when we dropped candidates that were not at least five years apart, the power estimates looked very similar. This result is promising, because we should expect a larger treatment effect size for this sample, given that there is more variation in age. When we dropped candidates that were less than 10 years apart, there is a slight decrease in power for any given treatment effect; since this group will likely have a larger average treatment effect, this comparison is about equally powerful as the other comparisons.

Given these results, we will focus on the t-tests and normal linear regression. These tests have the highest power when there is a large treatment effect, and they also do not wrongly identify treatment effects more than 5% of the time when no effect is present. Our primary analysis will look at candidates that were at least ten years apart, since the treatment effect should be larger for that subset, and there is not a large decrease in power or sample size

when we move from 5 years apart to 10 years apart. For secondary analyses we will examine five years and one year apart.

Ideally, we hope that the t-test and linear regression approaches will give us similar results. However, if there is a discrepancy between them, we will find what it is specifically that is making their results different.

This power analysis is based on two-sided tests of the null of zero effect ($\delta_0 = 0$), when the true effect is ($\delta_a = k$). In addition to testing the null of a zero effect, we are also interested in testing null hypotheses of large negative effects ($\delta_0 = -m$) and of large positive effects ($\delta_0 = m$), depending on the extent to which the literature offers reason to believe such effects are plausible. In most cases, the research community has substantial uncertainty about the sign and magnitude of causal effects and so doing so is informative. Testing a null of a large negative effect ($\delta_0 = -m$) against a true alternative of a positive effect ($\delta_a = m$) will have even more power, since the relevant comparison is greater ($\delta_a - \delta_0 = 2m$). In general we will report the confidence interval.

We may run additional tests using all the data that weight observations based on their age difference, since we expect units with larger age differences to have larger effects. One strategy for doing so is to run a separate analysis for each minimum age difference, and then combining them using Non-Parametric Combination.[5] This will lead candidates with small age differences to get less weight in our final hypothesis test, whereas candidates with large age differences would get higher weight.

7 Next Steps

We are also collecting data for other variables. In particular we have preliminary data for Left-Right status of the candidate and incumbency status of the party, which we intend to analyze in the above way.[6] Doing this analysis will help us determine whether to invest in coding the rest of the data for subsequent analysis. In any case we will publish the results from this examination.

References

[5]Caughey, Dafoe, Seawright. "Nonparametric Combination (NPC): A Framework for Testing Elaborate Theories." Journal of Politics.

[6]The Left-Right status will be coded on a three point scale, with 1 indicating that the winner was at least two points more to the right than the competitor (on the original five point scale), -1 if the winner is at least two points more to the left, and 0 otherwise. Change in incumbency status will be coded as -1, 0, or 1.