United We Fight: National Unity and Interstate Conflict

By

Andrew Douglas Bertoli

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy in Political Science in the Graduate Division of the University of California, Berkeley

Committee in Charge:

Professor Ron Hassner, Co-chair
Professor Jasjeet Sekhon, Co-chair
Professor Thad Dunning
Professor Aila Matanock
Professor Daniel Sargent

Summer 2016
Abstract

United We Fight: National United and Interstate Conflict

Andrew Douglas Bertoli

Doctor of Philosophy in Political Science

University of California, Berkeley

Professor Ron Hassner, Co-chair
Professor Jasjeet Sekhon, Co-chair

Does unity within countries increase conflict between them? In this dissertation, I investigate whether it does using natural experiments. The first essay tests how exogenous surges of nationalism affect state aggression by exploiting a natural experiment that was created by the World Cup. The second essay analyzes whether united government increases state aggression by looking at a similar natural experiment. The third essay considers some of the methodological challenges that arose over the course of this project and draws lessons for future research. Along with shedding new light on the relationship between domestic unity and interstate conflict, this project demonstrates the value of using natural experiments to study international security.
Introduction

A central question in the study of international relations is how domestic unity affects interstate conflict. While national cohesion is generally viewed as having many benefits at the domestic level, such as increasing government efficiency and economic growth, many scholars worry that it can have a negative impact on international relations. These scholars argue that two types of domestic unity are potentially dangerous. The first is social unity, or what is more commonly referred to as nationalism (Mearsheimer 1990). Among international relations scholars, nationalism is widely viewed as a major source of international conflict, and researchers have linked it to state aggression using a wide range of methodologies (Cederman, Warren, and Sornette 2011; Schrock-Jacobson 2012). The second is united government, which occurs when one party controls the main bodies of government. Many international relations scholars believe that united government increases the likelihood of interstate conflict, since it gives leaders more freedom to take aggressive action against other countries (Howell and Pevehouse 2005; Clark and Nordstrom 2006; Brulé 2006).

Nonetheless, determining how domestic unity affects interstate conflict is very difficult. First, the relationship may be explained by reverse causality. That is, countries might become more united in the lead-up to international conflict, which could make it appear that domestic unity has a large effect on state aggression when it really does not. Second, the relationship could be driven by confounding factors. For example, the underlying political disputes could increase national unity in the lead-up to conflict, inducing a correlation between domestic unity and state aggression that is not actually causal.

In this dissertation, I attempt to overcome these problems by using regression discontinuity designs. Regression discontinuity is a type of natural experiment that takes advantage of scoring systems with important cut-points. The classic example is a test where everyone who scores above a 50% receives a scholarship and everyone who scores below does not. If researchers wanted to analyze how winning the scholarship affected future income, they could compare the students who score 50’s to the students who score 49’s. The idea is that it should be nearly random which of them get the scholarship and which do not, so the data around the cut-point should be similar to experimental data.

The first paper in this dissertation uses a regression discontinuity involving the World Cup to test how nationalism affects state aggression. Specifically, I compare countries that barely qualified for the World Cup to countries that barely missed. I find that World Cup participation increases the likelihood that countries will initiate military disputes against other states. I also find that the countries that played against each other at that World Cup were much more likely to have military conflicts afterward. These results underscore the importance of being wary of other sources of nationalism, such as major national
achievements, tragedies, and revolutions. They also help legitimize concerns that the U.S. can provoke backlash by pursuing an activist grand strategy that foments nationalism in other countries.

The second paper in this dissertation tests how government unity affects state aggression. Specifically, I look at close national elections where parties barely achieved or barely fell short of united government. The results show that united government greatly increases the chances that democracies will initiate high-level disputes, and that it has no discernible impact on low-level disputes. The estimated effect is particularly large for the United States. These findings suggest that policymakers and the public should be alert to the potential danger of unnecessary conflicts when one party is in power. At the same time, they should be aware that divided governments might avoid some necessary conflicts, since opposition parties may wrongly oppose disputes where their countries should actually get involved. In short, believing that domestic politics “stop at the water’s edge” can increase the chances that important foreign policy mistakes go unnoticed.

The third paper discusses some of the methodological challenges that arose over the course of this project and draws lessons for future research. The first issue discussed arises when units are grouped into different strata. For example, in the scholarship test scenario, students might be grouped in to classes of different sizes, with the top five students from each class getting the scholarship. This complication can threaten the validity of the design if not handled properly. The second issue involves units that appear in the sample multiple times, which would occur if students were allowed to retake the exam in the test scholarship scenario. This issue can complicate the design, since some units would have more than one score. The third issue arises when the treatment that researchers care about is not actually assigned by the regression discontinuity. For example, a researcher might be interested in testing whether economic outcomes are better when Republicans or Democrats win close elections, even though the ideology of politicians is not as-if random. Researchers can learn important lessons from these cases, but they need to be careful when analyzing this type of design.

While these projects shed light on important substantive and methodological questions, they also demonstrate the promise of using natural experiments to study international conflict. The design-based inference movement that has led to breakthroughs in American and comparative politics is still just gaining traction in IR, due to the difficulty of studying major questions in international relations with natural experiments. This dissertation shows how important international relations topics can be addressed with design-based methods.
REFERENCES


Nationalism and Interstate Conflict: A Regression Discontinuity Analysis

Andrew Bertoli

27 March 2016

ABSTRACT. Does nationalism make interstate conflict much more likely, or are its effects largely constrained by strategic realities? This question matters for how scholars think and theorize about international relations, but answering it has been difficult because of endogeneity and measurement issues. To overcome these problems, I look to one of the most powerful sources of nationalism in the modern era—international sports. I begin by investigating several cases where surges of nationalism from sporting events led to military or political conflict between countries. I then analyze a regression discontinuity that was created by the format of the World Cup qualification process from 1958-2010. The results provide strong evidence that World Cup participation increases state aggression, especially for countries where soccer is the most popular sport. I also show that nationalism best explains these results by investigating two cases from the dataset and considering some additional evidence.

Scholars use nationalism to explain international conflicts from the Napoleonic Wars to the U.S. invasion of Iraq following September 11 (Rosato 2003; McCartney 2004; Cederman, Warren, and Sornette 2011). In fact, nationalism is one of the few factors that international relations theorists across paradigms view as an important force for interstate violence (Mearsheimer 1990; Wendt 1999; Keohane and Nye 2001). Scholars argue that it can increase enmity between countries (Walt 2011), undermine international cooperation (Rosato 2011), motivate societies to fight costly wars (Posen 1993), and cause governments to overestimate their relative military power (Snyder 2000). Moreover, many theories of war are premised on the idea that nationalism is an important source of conflict. Some examples include the theory that war can result from surges of nationalism following revolutions (Snyder and Mansfield 1995) and the theory that charismatic leaders who manipulate national sentiments make conflict more likely (Byman and Pollack 2001).

Nevertheless, whether fluctuations in nationalism can actually have a causal impact on interstate violence is a question that has thus far remained unanswered (Posen 1993; Schrock-Jacobson 2012). One concern is that national identities are already so salient that additional increases in nationalism may have a very small effect on the likelihood of war. Another is that surges of nationalism might make little difference given the constraints
imposed on states by the international system. Past studies that have sought to address this question have faced two major challenges. First, there is no comprehensive dataset that tracks nationalism across the international system, or even an established way to measure it in any given case. Second, even though nationalism often appears to increase in the lead-up to war, it is very difficult to show that this relationship is causal. For instance, the existing political disputes underlying the conflict might create the nationalism, or leaders might try to incite it to increase morale for a future war (Posen 1993).

This paper tests how fluctuations in nationalism influence state aggression by looking at surges of nationalism that were created by international sports. Sporting events are an ideal place to look because it is widely accepted that they increase the salience of national identities. Scholars have shown that they often lead to feelings of national unity and hatred towards other countries (Vincent et al. 2010). These sentiments have resulted in frequent outbreaks of nationalistic violence at games (Markovits and Rensmann 2010), and at times have intensified rivalries at the international level (Reid 1999; Hitchens 2010; Schama 2014). Some famous examples include the 1969 Football War between El Salvador and Honduras, the 2009 Egypt-Algeria World Cup Dispute, the 1936 Nazi Olympics, and the controversial Moscow Dynamo soccer club trip to Britain in 1945 (Orwell 1945; Lindsey 2009).

To overcome the endogeneity problem, I analyze a regression discontinuity that was created by the format of the World Cup qualification process from 1958-2010. Over this period, many countries qualified for the World Cup by playing a round of games against other states and achieving a top position in the final standings. It is therefore possible to compare the group of countries that barely qualified to the group that barely missed. These countries went to the World Cup or stayed home based on small differences in their records after many games, so the data obtained by this procedure should be similar to what would be expected in a randomized experiment.

Using this approach, I construct a sample of countries that barely qualified or barely missed the World Cup. Specifically, I select pairs of countries where one country qualified and the other fell short, provided that they were separated by no more than two points in the standings and the qualifier scored at least five points. I made these design choices prior to collecting the data, believing that they would lead to a sufficiently large sample under which qualification was as-if random. In total, the sample consists of 142 countries, with 71 barely qualifying and 71 barely missing. The qualifiers and non-qualifiers are balanced across a wide range of political, economic, and demographic factors, which supports the as-if random assumption. They are also balanced on past levels of aggression, which I measure in the standard way as the number of militarized interstate disputes that a country initiates.

The results show that going to the World Cup increases aggression substantially. The countries that barely qualified experienced a significant spike in aggression during the World Cup year. The disputes that they started also tended to be much more violent.
than the disputes started by the non-qualifiers. The results hold under various robustness checks, and the estimated treatment effect is much higher for countries where soccer is the most popular sport. Substantively, the estimates suggest that going to the World Cup increases state aggression by about two-fifths as much as a revolution does, and that it has roughly the same effect as electing a leader who is backed by the military.

It is also possible to replicate the analysis using the FIFA regional soccer championships like the European Football Championship and the African Cup of Nations. In total, this new sample consists of 78 countries that barely made or barely missed their regional soccer tournaments. The qualifiers and non-qualifiers were again well-balanced on aggression levels prior to qualification, but the qualifiers became significantly more aggressive following qualification. The results from this analysis are available in the Online Appendix.

This paper proceeds as follows: I first review the existing theoretical and empirical work about the relationship between nationalism and war. I then elaborate on why international sports provide an excellent opportunity to test whether exogenous surges of nationalism have a major causal impact on state aggression. Next, I examine some qualitative evidence from the history of international sports that suggests that nationalism can lead to conflict on the international stage, and I consider some of the mechanisms through which this process can unfold. I then explain my research design in more depth and present the results from the World Cup analysis. After that, I explore two cases from the dataset and perform some additional tests. The final section concludes.

SECTION 1: EXISTING LITERATURE ON NATIONALISM AND WAR

Nationalism is generally defined as the practice of identifying with a nation-state and viewing other nations as competitors that are culturally or morally inferior (Orwell 1945). For many, this identification is so powerful that it compels them to kill and die for their countries (Rosato 2003). While this devotion may be somewhat difficult to explain at the individual level, most scholars agree that it is a widespread phenomenon (Anderson 1983; Robinson 2014), and that it plays a central role in international relations (Mearsheimer 1990; Wendt 1999). Most importantly, it defines the boundaries of the international system that scholars often take for granted. It also helps explain how major war between countries is possible by ensuring that there will be people who are willing to die for the national interest.

Nevertheless, there is much less certainty about whether surges of nationalism have an important causal impact on state behavior. The effects of nationalism might be greatly constrained by strategic factors, or nationalism might not fluctuate enough to have noticeable effects at the international level. A central problem in answering this question has been endogeneity. As Posen (1993) explains, “Leaders use nationalism to mobilize public support for military preparation and sacrifices. When war seems imminent, for any reason, the intensity of propaganda increases. The same is true when wars last for any length of
time. Thus it will often be difficult to show that nationalism caused any conflict, because it will generally be accompanied and accentuated by other causes of the conflict.”

But despite this concern, many international relations scholars believe that nationalism has been an important source of conflict throughout history. For instance, they have used it to explain the conquests of Imperial Germany and Japan, the breakup of the Soviet Union, and the Balkans Wars (Mearsheimer 1990; Snyder and Mansfield 1995; Van Evera 1994). The view that nationalism can be a force for conflict also underpins many important arguments made by international relations scholars, like that bombing civilians during wartime makes the target population less willing to surrender (Pape 1996) and that the United States can provoke backlash by stationing military troops abroad (Posen 2013). No doubt, the idea that nationalism can cause interstate violence influences how many scholars think about international relations.

Moreover, many theories of war are predicated on the idea that high levels of nationalism make conflict more likely. For instance, this idea is central to the theory that emerging democracies are likely to start wars as a result of the hypernationalistic rhetoric of their leaders who seek to increase their domestic support (Snyder and Mansfield 1995; Snyder 2000). Similarly, this idea underpins the theory that war can be caused by charismatic leaders who have the ability to rally mass nationalism (Byman and Pollack 2001). It is also key to theories that focus on variation in nationalism across the international system. For example, the idea that grievances from past wars or atrocities can explain why some nations fight each other is based on the idea that the intensity of nationalism can have an important impact on the likelihood of conflict (Van Evera 1994).

In short, knowing the extent to which nationalism affects state aggression matters for how scholars think and theorize about international relations. While past quantitative efforts to investigate this question have remained limited, in part because there is no comprehensive dataset that tracks nationalism, a number of recent studies have made noteworthy progress. First, several experiments have found that people tend to think more hawkishly when they see their national flags (Hassin 2007; Kemmelmeier and Winter 2008). These studies provide important micro-level evidence of a causal link between nationalism and conflict. However, they fall short of showing that nationalism affects state behavior, because they only look at the responses of individuals.

Second, Schrock-Jacobson (2012) recently provided a major contribution to this research program by conducting the first large-$N$ cross-national study that tests the relationship between nationalism and military conflict directly. To deal with the measurement problem, she selected a random sample of state-years from 1816-1997 and coded for whether countries experienced a nationalistic movement in those years. Using regression, she shows that there is a correlation between nationalism and the onset of war that does not disappear after controlling for some baseline factors. While this result provides valuable observational evidence for a causal relationship between nationalism and conflict, it may
be explained by omitted variable bias. This problem is concerning because countries usually become more nationalistic in the lead-up to war, and the anticipation of future conflict is a difficult variable to observe or proxy.

Thus, measurement and endogeneity issues have made it difficult to test whether surges of nationalism can have an important impact on interstate violence. In the next section, I propose a new strategy to overcome these challenges.

SECTION 2: USING INTERNATIONAL SPORTS AS A SOURCE OF NATIONALISM

International sports provide an excellent opportunity to test whether nationalism has a causal effect on interstate conflict. First, they offer a way around the measurement problem. While past studies have struggled to find a reliable, comprehensive measure of nationalism that spans across countries and time, most political scientists agree that international sporting events increase it (Cha 2009; Markovits and Rensmann 2010; Walt 2011). Sociologists have also shown that international sports often make the national discourses within countries more hawkish, with reporters describing games in military terminology and comparing wins and losses to past battles (Garland and Rowe 1999). As Vincent (2010) explains, “only warfare feeds the imagination and cements national identity more than sports.”

There are many examples where international sports have increased national unity within countries. One of the most famous cases was Nelson Mandela’s use of rugby to unite South Africa following the end of the apartheid regime (Steenveld and Strelitz 1998). Similarly, qualification for the 2006 World Cup helped unify the Ivory Coast after four years of civil war (Mehler 2008). Government leaders in Yemen also used the national soccer team to help break down divisions between the northern and southern regions of the country following integration in 1990 (Cha 2009). In short, international sports can evoke feelings of nationalism that spill over into the political realm.

A second advantage is that it is very hard to imagine how sports could increase the chances of interstate conflict in any way besides generating nationalism. The most plausible alternative explanation may be that sports distract the public, giving leaders more freedom to do what they want on the international stage. However, this mechanism would only explain a very short term effect, which is inconsistent with the data that I present later in this paper. Other possibilities that are unrelated to nationalism are hard to think of, in part because international sports and nationalism are inextricably linked. The vast majority of fans root for their countries because of national identification, and the very act of cheering for one’s side is an expression of nationalism (Hobsbawm 1990).

A third major advantage of using international sports is that the most popular sporting event, the World Cup, created a regression discontinuity where surges of nationalism were essentially exogenous. As I will discuss more in the coming sections, qualification for the
World Cup was as-if random for a large number of countries. This natural experiment allows us to avoid the endogeneity problem that has left past research somewhat inconclusive.

There are also many historical cases that suggest that international sports increase the chances of interstate violence by inciting nationalism, so it is plausible that this design could work. I will turn to these cases in the next section to provide some preliminary qualitative evidence that nationalism from sporting events can increase state aggression and to consider some of the ways that this process can occur.

SECTION 3: INVESTIGATING THE HISTORICAL EVIDENCE

The idea that nationalism from sporting events can increase the likelihood of interstate conflict has been advanced by many writers (Orwell 1945; Reid 2000; Hitchens 2010; Markovits and Rensmann 2010; Schama 2014). These scholars have identified a number of cases where nationalism from sporting events appears to have led to major military or political conflict between countries. Their research suggests that conflict is often the unintended consequence of sports nationalism, but that leaders also sometimes exploit international sporting events to generate support for their foreign policy ambitions. In either case, mass nationalism increases leaders’ opportunity and incentive to go on the offensive. I discuss these cases below, and then I conclude by highlighting some important points to note from history.

1969 Football War. Many studies have concluded that this war was caused by a series of soccer riots that exacerbated an already tense political situation between El Salvador and Honduras (Anderson 1981; Reid 2000; Kapuscinski 2013). In the late 1960s, the Honduran government began taking land from the Salvadoran immigrants living in its country and giving it to Honduran-born peasants in response to unrest among Honduran agricultural workers. This policy led to skirmishes between Honduran and Salvadoran nationals and prompted a media war between the two countries that further encouraged xenophobia.

In this volatile political atmosphere, Honduras and El Salvador played three World Cup qualification games. The night before the first game in Honduras, a mob of Hondurans surrounded the hotel where the Salvadoran players were staying and made as much noise as possible to prevent them from sleeping. The next day, El Salvador lost 1-0, but stories of the incident were widely reported throughout El Salvador and incited public outrage. When the Honduran team traveled to El Salvador to play the next game, Salvadoran fans threw rocks through the windows of the hotel where the Honduran players were staying and used drums, horns, and bands to keep the team up all night. The crowd was so hostile that the next morning the Honduran players had to be driven to and from the stadium in armored cars, and the Salvadoran army was needed to provide security at the game. Outside the stadium, Salvadoran fans assaulted many Hondurans, burned their cars, and in some cases chased them to the border. Honduras lost the game 3-0.
When news of what happened reached Honduras, it prompted widespread violence against Salvadoran nationals living in the country, and thousands were forced to flee to El Salvador. Tensions escalated again when the countries played a tie-breaking game thirteen days later in Mexico City that needed to be supervised by the Mexican military. That day, the two sides broke diplomatic ties and El Salvador formally accused the Honduran government of human rights violations for not protecting the Salvadoran nationals living in its country. Less than three weeks later, El Salvador invaded Honduras, starting a war that caused about two thousand casualties (Reid 2000).

**2009 Egyptian-Algerian World Cup Dispute.** This conflict began after Egyptian fans attacked a bus transporting the Algerian team and injured three players. Algerians retaliated against Egyptian companies and property in Algeria. In response, Mubarak removed the Egyptian ambassador from Algiers, gave an impassioned speech about Egypt’s insulted national pride, and accused the Algerian government of staging the attacks on its players (Lindsey 2009). Egypt also sent a plane into Algeria to rescue Egyptian nationals from the soccer violence, although the aircraft was not granted permission to land (Shenker 2009). As Hitchens (2010) describes, “Before the match in Khartoum, Egypt and Algeria had no diplomatic quarrel. After the game, perfectly serious people in Cairo were saying the atmosphere resembled that following the country’s defeat in the June 1967 war.” In fact, the Arab League responded to this incident by proposing that politicians from its member states should not attend sensitive sports games because these events could make them more hostile toward other countries (Belmary 2009).

**2014 Serbian-Albanian Drone Conflict.** Serbian and Albanian soccer fans have a history that is so marred with violence that they are not allowed to attend each other’s games. At a UEFA Cup qualifying game in Serbia on October 14, 2014, Albanian fans flew a small drone into the stadium that was carrying an Albanian flag. When the drone got close enough to the ground, a Serbian player ripped the flag off, sparking a fight between the Serbian and Albanian teams. Serbian fans rushed the field to attack the Albanian players, who were barely able to get back to their locker room.¹ Serbia called the drone stunt an “act of terrorism” designed to provoke Serbian fans and claimed that the brother of the Albanian president was behind it (Riach 2014). Albania accused Serbian police of attacking Albanian players.

**1934 Italian World Cup.** Throughout the 1930s, Mussolini intentionally used sports to increase support for the Italian war machine. As Tunis (1936) describes, the main purpose of sports for Mussolini was “the mass production of cannon fodder.” He writes, “sport ceased to be a free activity and became a function of the government... The results are watched, collected, catalogued and exploited, at home and abroad.” Similarly, Martin

---

¹. A video of this incident is available at [http://www.tubchop.com/watch/6035427](http://www.tubchop.com/watch/6035427).
Table 1  Notable Cases Where Scholars Have Claimed That Sports Nationalism Led to Interstate Conflict

1. Football War (1969)
2. Egypt-Algeria World Cup Dispute (2009)
4. Italian World Cup (1934)
5. Nazi Olympics (1936)
6. Bodyline (1932)
7. Moscow Dynamo Soccer Trip to Britain (1945)

(2004) explains that the 1934 World Cup “was more like a fascist rally than a sporting contest.” Goldblatt (2008) draws a clear link between this competition and Italian aggression toward other countries. As he describes, “The preparations for the tournament coincided with a steadily more expansionist and aggressive foreign policy that would culminate after the World Cup in the invasion of Abyssinia (Ethiopia), intervention in the Spanish Civil War and relentless pressure on Albania and Central Europe.”

1936 Nazi Olympics. Following Mussolini’s example, Hitler used the 1936 Nazi Olympics to provoke feelings of Aryan supremacy and victimization in Germany during the lead-up to World War II (Rippon 2006). Large (2007) describes the games as “a crucial part of the Nazi regime’s ‘spiritual mobilization’ to win the hearts and minds of the German people,” and Schama (2014) argues that they played a pivotal role in creating domestic support for German expansion. In fact, Hitler did not care about sports until his advisers convinced him that they could be used to generate support for the German war machine (Krüger 1998). Of course, these scholars are not suggesting that World War II would not have happened without the 1936 Olympics, but the case illustrates how leaders with military ambitions can use nationalism from sports to increase public support for aggression.

Other Cases. Table 1 lists the examples discussed above along with some other notable cases where scholars have argued that nationalism from sporting events led to major military or political conflict between countries. Bodyline (1932) was a dispute between Britain and Australia over cricket that resulted in rioting, nationalistic vandalism, boycotts, and other economic fallout. Many Australians consider Bodyline to be one of the two main reasons that Anglo-Australian relations collapsed in the 1930s, with the other being the Great Depression (Frith 2013; Swan 2013). The Moscow Dynamo soccer trip to Britain
(1945) involved a series of games that were supposed to help strengthen relations between Britain and the Soviet Union following World War II. However, they led to fights and numerous controversies over cheating, and Britain canceled the tour early when it became clear that the Soviets were exploiting the games to generate nationalism at home (Orwell 1945; Kowalski and Porter 1997). I also include the Croatian War of Independence as a case because it was prefaced by a substantial amount of soccer violence. In fact, many Serbs and Croats consider a battle in the Croatian soccer stadium to be the unofficial beginning of the war (Đorđević 2009). While there were major underlying political disputes at the time, Sack and Suster (2000) argue that “it would be a mistake to view these matches as mere epiphenomena mirroring larger social and political events but having no power to influence them.”

**Key Points.** These cases illustrate three key points about conflicts from international sporting events:

1. **The nationalism can be unintended or driven by government leaders.** In many cases, especially the Football War and Bodyline, international tension arose simply from the passions that sports evoke. Emotions over the sporting event changed public opinion, and leaders then had to respond. However, in other cases, most notably the 1934 World Cup and the 1936 Olympics, leaders deliberately exploited international sports to generate public support for their aggressive foreign policy agendas. The sporting event gave them a way to create a surge of nationalism in their country that could change public opinion. Thus, these cases suggest that nationalistic conflict may be accidental or elite-driven. Both causal paths appear to be valid ways that nationalism can intensify international rivalry.

2. **The conflict is sometimes between countries that compete against each other head-to-head, but not always.** As the 1934 World Cup and the 1936 Olympics demonstrate, nationalism from sporting events can increase the likelihood of conflict between countries even when they do not play against each other directly. Ethiopia did not attend the 1934 World Cup, but it was the first target of Italian expansion in 1935. Similarly, the Soviet Union and Jewish community boycotted the 1936 Olympics, although they found themselves the victims of German nationalism in the coming years. This idea that sports nationalism can lead to conflict indirectly is also supported by Iraq’s experience at the 2007 Asian Cup, which Stephens (2007) claims increased Iraqi nationalism and feelings of resentment toward the United States. Similarly, when countries play games against states that they do not have an international rivalry with, it is common for their fans to burn flags of their traditional rival countries.

   This distinction between the direct and indirect effects of sports nationalism is important because the main test in this paper focuses on the indirect effect. Since countries play only a few other states at the World Cup, most of whom are far away geographically, the test in this paper primarily relies on indirect conflicts. Thus, it investigates whether
state aggression can increase from general nationalism (like in the case of the French Revolution), rather than from nationalism that is directed at a specific rival country (like in the case of the Football War). However, I also show in the results section that the pairs of countries that played against each other at the World Cup became much more likely to engage in military disputes afterward.

(3) Nationalism from sports appears capable of having downstream effects. In cases like the 1934 World Cup and Croatian War of Independence, tension from sports did not immediately translate into conflict at the international level. Rather, it appears to have primed people to think more aggressively towards other countries, thereby influencing their behavior when disputes did arise in the future. The analog case in U.S. history might be how nationalism from September 11 made many Americans much more willing to invade Iraq over a year-and-a-half later (McCartney 2004). Thus, we should expect that the World Cup may have downstream effects that appear a year or two afterward.

SECTION 4: DESIGN

Strategy for Identifying the Causal Effect. I estimate the impact of World Cup participation on state aggression by using a regression discontinuity design. This method has become an increasingly popular research technique in the social sciences over the last decade for its ability to identify causal effects (Dunning 2012). It can be used when a treatment is given to units that surpass a certain cut-point in a scoring system. The idea is to compare the group of units that scored just above the cut-point to the group that scored just below it. For example, if there was a test where everyone who scored a 50% or higher received a scholarship, we could assess the impact of winning the scholarship by comparing people who scored 50% to people who scored 49%. As long as there is some randomness in the scoring process, it should be close to random who ended up on either side of the cut-point (Lee 2008). Thus, this method should produce the type of data that we would expect in a randomized experiment.

There are two main approaches to analyzing regression discontinuities. The first is to compare the units that scored just above the cut-point to units that scored just below it using normal experimental tests (Dunning 2012; Cattaneo, Frandsen, and Titiunik 2015). The advantage of this approach is that it makes very few statistical assumptions, since it is essentially a natural experiment with a quasi-random treatment. The major decision that researchers must make is how to define barely making and barely missing the treatment. In the test example, researchers might look at the students who scored 49% and 50%, or they might also want to include students who scored 48% and 51%.

The second approach is to estimate the difference in the potential outcomes at the cut-point using two regression lines (Voeten 2013). This approach does not assume as-if randomness in a small window around the cut-point. Instead, it requires that the potential
outcomes are smooth in an area around the cut-point and that they can be estimated reasonably well using regression. Researchers using this approach must make a number of decisions, including which smoother and bandwidth selection procedure to use and how to estimate the confidence intervals for the regression lines.

My results are significant for both approaches, although I focus on the first approach in this paper. The main reason is that the standard procedures for the second approach are not valid when dealing with a scoring system that has discrete values, such as the World Cup qualification process (Lee and Card 2008; Cattaneo, Frandsen, and Titiunik 2015). In addition, the first approach is more intuitive and requires less design choices. It also does not depend on regression predictions at the cut-point, which may be unreliable since the cut-point is an endpoint for both regression lines.

To address the fact that which countries barely qualify and barely miss is not perfectly random, I use a difference-in-differences estimator to reduce any possible bias. Thus, I compare the change in aggression for the countries that barely qualified to the change in aggression for the countries that barely missed. As a robustness check, I also control for baseline differences between the qualifier and non-qualifier groups using linear regression. The findings remain significant under all specifications, which are primarily detailed in the supporting information. I also show that the results are insensitive to the limited number of design choices that I made, such as how I defined the bare qualifiers and bare non-qualifiers. I explain these choices in the next section.

**Constructing the Treatment and Control Groups.** The World Cup takes place once every four years during the summer, with the qualification process ending the winter before. The first qualification round was held in 1934, and it has been in place ever since. The host country qualifies automatically, as did the winner of the previous World Cup up until 2010. All other countries are required to play their way in. They do so in one of two ways: (1) by playing a round of games against other states in their region and earning a certain position in the standings or (2) by winning a play-in game or several playoff games. Either way, the format is set well in advance.

This study focuses on the standings format, which is illustrated in Table 2. In the standings format, countries that were very close to the qualification cut-point went to the World Cup or stayed home based on small differences in their records after many games. Thus, qualification should be close to random for these countries. This as-if randomness should make the group of states that barely qualified similar to the group that barely missed across observable and unobservable factors. Moreover, we can check that the qualifiers and non-qualifiers are balanced to verify that the design worked.

The playoff format is less easy to exploit with regression discontinuity analysis. Countries make or miss the World Cup based on their performance in the final round, which

---

2. Before 1994, countries earned two points for a win, one point for a tie, and nothing for a loss. Starting in 1994, the value of a win was increased to three points.
Table 2  Example of the Final Standings from a 1994 Qualification Round

<table>
<thead>
<tr>
<th>Rank</th>
<th>Country</th>
<th>Score</th>
<th>Qualified</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Italy</td>
<td>16</td>
<td>Yes</td>
</tr>
<tr>
<td>2</td>
<td>Switzerland</td>
<td>15</td>
<td>Yes</td>
</tr>
<tr>
<td>3</td>
<td>Portugal</td>
<td>14</td>
<td>No</td>
</tr>
<tr>
<td>4</td>
<td>Scotland</td>
<td>11</td>
<td>No</td>
</tr>
<tr>
<td>5</td>
<td>Malta</td>
<td>4</td>
<td>No</td>
</tr>
<tr>
<td>6</td>
<td>Estonia</td>
<td>1</td>
<td>No</td>
</tr>
</tbody>
</table>

Note: The sample consists of pairs of countries like Switzerland and Portugal that barely made and barely missed qualification.

could be correlated with other factors that are related to their likelihood of future aggression. This problem is particularly concerning because the last game in the playoff format is often played between countries from different regions that have large disparities in terms of GDP and population, along with many other factors. However, the findings in this study remain significant when close playoff games are included, so the results are not sensitive to whether the outcomes of these games are considered random or not.

Using data from the standings format from 1934-2010, I select pairs of countries that were separated by no more than two points in the standings, provided that the winner scored at least five points. In total, the sample consists of 142 countries, which are listed in Table 3. There were seven pairs before 1958 where the winner scored less than five points. In these cases, the teams played only a small number games, typically three or less each. Since the goal of the design was to obtain two groups of countries that were assigned to treatment or control based on small differences in their records after many games, these countries were dropped from the analysis.

Seventeen of the 71 pairs tied in the standings. Nine of these ties were broken by a playoff game, seven were decided by looking at which country had the larger average margin of victory, and one was decided by which team scored more goals. I include these pairs in the analysis for two reasons. First, the playoff games were more like toss-ups because they were played between teams of comparable skill. Second, there are not strong reasons to believe that comparable teams would sort based on margin of victory or total goals scored. Nevertheless, the results remain significant whether ties are included or not.

Lastly, I had to remove pairs where one of the teams did not represent a country. I exclude Scotland, Northern Ireland, and Wales from this analysis, and I count England as Britain. I also exclude the Representation of Czechs and Slovaks, which was a union.
Table 3  Countries That Barely Made and Barely Missed the World Cup

<table>
<thead>
<tr>
<th>Qualifier</th>
<th>Non-qualifier</th>
<th>Year</th>
<th>Qualifier</th>
<th>Non-qualifier</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yugoslavia</td>
<td>Romania</td>
<td>1958</td>
<td>Tunisia</td>
<td>Egypt</td>
<td>1978</td>
</tr>
<tr>
<td>France</td>
<td>Belgium</td>
<td>1958</td>
<td>France</td>
<td>Ireland</td>
<td>1982</td>
</tr>
<tr>
<td>Austria</td>
<td>Netherlands</td>
<td>1958</td>
<td>Austria</td>
<td>Bulgaria</td>
<td>1982</td>
</tr>
<tr>
<td>Soviet Union</td>
<td>Poland</td>
<td>1958</td>
<td>Britain</td>
<td>Romania</td>
<td>1982</td>
</tr>
<tr>
<td>Hungary</td>
<td>Bulgaria</td>
<td>1958</td>
<td>Peru</td>
<td>Uruguay</td>
<td>1982</td>
</tr>
<tr>
<td>Britain</td>
<td>Ireland</td>
<td>1958</td>
<td>El Salvador</td>
<td>Mexico</td>
<td>1982</td>
</tr>
<tr>
<td>Paraguay</td>
<td>Uruguay</td>
<td>1958</td>
<td>New Zealand</td>
<td>China</td>
<td>1982</td>
</tr>
<tr>
<td>Argentina</td>
<td>Bolivia</td>
<td>1958</td>
<td>Portugal</td>
<td>Sweden</td>
<td>1986</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>France</td>
<td>1962</td>
<td>Soviet Union</td>
<td>Switzerland</td>
<td>1986</td>
</tr>
<tr>
<td>Switzerland</td>
<td>Sweden</td>
<td>1962</td>
<td>Bulgaria</td>
<td>East Germany</td>
<td>1986</td>
</tr>
<tr>
<td>Portugal</td>
<td>Czechoslovakia</td>
<td>1966</td>
<td>Romania</td>
<td>Denmark</td>
<td>1990</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>Belgium</td>
<td>1966</td>
<td>Austria</td>
<td>Turkey</td>
<td>1990</td>
</tr>
<tr>
<td>West Germany</td>
<td>Sweden</td>
<td>1966</td>
<td>Czechoslovakia</td>
<td>Portugal</td>
<td>1990</td>
</tr>
<tr>
<td>Chile</td>
<td>Ecuador</td>
<td>1966</td>
<td>United States</td>
<td>Trinidad</td>
<td>1990</td>
</tr>
<tr>
<td>Czechoslovakia</td>
<td>Hungary</td>
<td>1970</td>
<td>UAE</td>
<td>Qatar</td>
<td>1990</td>
</tr>
<tr>
<td>Romania</td>
<td>Greece</td>
<td>1970</td>
<td>Ireland</td>
<td>Denmark</td>
<td>1994</td>
</tr>
<tr>
<td>Bulgaria</td>
<td>Poland</td>
<td>1970</td>
<td>Switzerland</td>
<td>Portugal</td>
<td>1994</td>
</tr>
<tr>
<td>Italy</td>
<td>East Germany</td>
<td>1970</td>
<td>Bulgaria</td>
<td>France</td>
<td>1994</td>
</tr>
<tr>
<td>Sweden</td>
<td>France</td>
<td>1970</td>
<td>Netherlands</td>
<td>Britain</td>
<td>1994</td>
</tr>
<tr>
<td>Belgium</td>
<td>Yugoslavia</td>
<td>1970</td>
<td>Bolivia</td>
<td>Uruguay</td>
<td>1994</td>
</tr>
<tr>
<td>Peru</td>
<td>Bolivia</td>
<td>1970</td>
<td>Cameroon</td>
<td>Zimbabwe</td>
<td>1994</td>
</tr>
<tr>
<td>Morocco</td>
<td>Nigeria</td>
<td>1970</td>
<td>Nigeria</td>
<td>Ivory Coast</td>
<td>1994</td>
</tr>
<tr>
<td>Sweden</td>
<td>Austria</td>
<td>1974</td>
<td>Morocco</td>
<td>Zambia</td>
<td>1994</td>
</tr>
<tr>
<td>Netherlands</td>
<td>Belgium</td>
<td>1974</td>
<td>South Korea</td>
<td>Japan</td>
<td>1994</td>
</tr>
<tr>
<td>East Germany</td>
<td>Romania</td>
<td>1974</td>
<td>Jamaica</td>
<td>Costa Rica</td>
<td>1998</td>
</tr>
<tr>
<td>Poland</td>
<td>Britain</td>
<td>1974</td>
<td>Chile</td>
<td>Peru</td>
<td>1998</td>
</tr>
<tr>
<td>Uruguay</td>
<td>Colombia</td>
<td>1974</td>
<td>Senegal</td>
<td>Morocco</td>
<td>2002</td>
</tr>
<tr>
<td>Argentina</td>
<td>Paraguay</td>
<td>1974</td>
<td>Nigeria</td>
<td>Liberia</td>
<td>2002</td>
</tr>
<tr>
<td>Haiti</td>
<td>Trinidad</td>
<td>1974</td>
<td>Ivory Coast</td>
<td>Cameroon</td>
<td>2006</td>
</tr>
<tr>
<td>Italy</td>
<td>Britain</td>
<td>1978</td>
<td>Tunisia</td>
<td>Morocco</td>
<td>2006</td>
</tr>
<tr>
<td>Austria</td>
<td>East Germany</td>
<td>1978</td>
<td>Togo</td>
<td>Senegal</td>
<td>2006</td>
</tr>
<tr>
<td>France</td>
<td>Bulgaria</td>
<td>1978</td>
<td>Angola</td>
<td>Nigeria</td>
<td>2006</td>
</tr>
<tr>
<td>Poland</td>
<td>Portugal</td>
<td>1978</td>
<td>Algeria</td>
<td>Egypt</td>
<td>2010</td>
</tr>
<tr>
<td>Sweden</td>
<td>Norway</td>
<td>1978</td>
<td>Nigeria</td>
<td>Tunisia</td>
<td>2010</td>
</tr>
<tr>
<td>Spain</td>
<td>Romania</td>
<td>1978</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
of players from the Czech Republic and Slovakia that played from 1992-93. No other changes were necessary.

**Measuring Aggression.** Similar to past studies (Lai and Slater 2006; Colgan 2010), I measure aggression using the number of militarized interstate disputes (MIDs) that a country initiates. These disputes are instances where states explicitly threaten, display, or use force against other countries (Ghosn, Palmer, Bremer 2004). This measure is commonly used in security studies, since wars happen too infrequently to be a useful measure in most statistical tests. While most MIDs are low-level and not very interesting in their own right, they are a good proxy for the likelihood of major interstate conflict. They are indicative of the foreign policy stance of a country and whether that country is willing to initiate conflicts with its rivals. I also show in the next section that the results remain significant for high-level disputes that involved a direct attack, clash, or the start of interstate war.

**Checking for Balance.** The goal of the design was to achieve balance across observable and unobservable pre-treatment characteristics. The qualifier and non-qualifier groups should be similar except that the qualifiers went to the World Cup and the non-qualifiers did not.

Of course, we should not expect the two groups to look exactly the same on pre-treatment factors. There will be some baseline differences simply by chance. In randomized experiments, the p-values for pre-treatment characteristics should be distributed uniformly between 0 and 1. Thus, there are statistically significant differences at the 5% level for about one out of every 20 pre-treatment characteristics, simply because of chance variation. But the inferences drawn from this study will be most credible if the qualifiers and non-qualifiers are balanced on all observable characteristics that might influence future aggression, as well as on past levels of aggression.

Figure 1 provides a comparison between the two groups. Since the qualification round ends about six months prior to the summer of the World Cup, the data used in this summary were taken from the year before the World Cup. The qualification rounds finished at the end of this year. For each variable, I plot the p-value from a two-tailed paired t-test that evaluates the difference in means between the qualifier and non-qualifier groups. I also plot the p-values for aggression levels in the years leading up to the World Cup.

Overall, the balance between the qualifiers and non-qualifiers looks similar to what would be expected in a randomized experiment. The p-values seem to be distributed uniformly between 0 and 1, which supports the idea that this data is like experimental data. Most importantly, the qualifiers and non-qualifiers are balanced on levels of aggression in the years leading up to the World Cup. In fact, this comparability extends much further back. If you compare the aggression levels of the qualifiers and non-qualifiers in each of the 50 years prior to qualification, there is statistical significance only twice at the 5% level, in Year -10 and Year -36. Aside from these years, the difference in aggression levels between the two groups is never significant at even the 10% level. The p-value over the
Figure 1  Balance Between the Qualifiers and Non-Qualifiers

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Treatment Mean</th>
<th>Control Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Population</td>
<td>33,438,500</td>
<td>35,742,500</td>
</tr>
<tr>
<td>Urban Population</td>
<td>9,189,200</td>
<td>8,879,540</td>
</tr>
<tr>
<td>Imports</td>
<td>23,614,700,000</td>
<td>18,628,100,000</td>
</tr>
<tr>
<td>Exports</td>
<td>21,695,600,000</td>
<td>18,894,900,000</td>
</tr>
<tr>
<td>Material Power Score</td>
<td>0.01</td>
<td>0.009</td>
</tr>
<tr>
<td>Level of Democracy</td>
<td>0.465</td>
<td>0.465</td>
</tr>
<tr>
<td>Engaged in Civil War</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>Resolved Civil War</td>
<td>0.01</td>
<td>0</td>
</tr>
<tr>
<td>Year of State Formation</td>
<td>1881</td>
<td>1877</td>
</tr>
<tr>
<td>Birth Rate</td>
<td>23.6</td>
<td>24</td>
</tr>
<tr>
<td>Infant Mortality</td>
<td>51.1</td>
<td>50.6</td>
</tr>
<tr>
<td>Life Expectancy</td>
<td>65.6</td>
<td>66</td>
</tr>
<tr>
<td>Median Age</td>
<td>27.7</td>
<td>27.7</td>
</tr>
<tr>
<td>Number of Alliances</td>
<td>15.7</td>
<td>16.5</td>
</tr>
<tr>
<td>U.S. Ally</td>
<td>0.4</td>
<td>0.479</td>
</tr>
<tr>
<td>Soccer Most Popular Sport</td>
<td>0.93</td>
<td>0.93</td>
</tr>
<tr>
<td>Appearance at Previous World Cup</td>
<td>0.352</td>
<td>0.31</td>
</tr>
<tr>
<td>MIDs Initiated in the Year Before</td>
<td>0.183</td>
<td>0.113</td>
</tr>
<tr>
<td>MIDs Initiated in the 3 Years Before</td>
<td>0.69</td>
<td>0.577</td>
</tr>
<tr>
<td>MIDs Initiated in the 5 Years Before</td>
<td>1.07</td>
<td>0.873</td>
</tr>
</tbody>
</table>

entire 50-year period is $p = 0.46$, with the qualifiers averaging 0.04 disputes per year more than the non-qualifiers.

Thus, the data passes an important placebo test. It is unlikely that the balance plot left out some factor that would make the qualifiers significantly more aggressive than the non-qualifiers after qualification, since this factor should have also made the qualifiers much more aggressive before qualification. So aside from the treatment effect, there is little reason to suspect that the qualifiers would behave much more aggressively than the non-qualifiers after qualification.

SECTION 5: FINDINGS

Although the qualifiers and non-qualifiers were comparable on past levels of aggression, the story is much different following qualification. Figure 2 tracks the aggression levels of the two groups over this period. Prior to qualification, they had very similar records
of aggression. However, the qualifiers became much more aggressive following qualification, and they remained so until about the second year after the tournament. The fact that the increase in aggression begins after qualification is consistent with historical evidence showing that many countries experience a surge of nationalism when their countries qualify for the World Cup (Ralph 2007; Mehler 2008).

Note that the aggression levels of the two groups drop the year before qualification, and then the qualifiers spike while the non-qualifiers stay low. This trend is consistent with the theory that the World Cup causes conflict. Recall that four years before qualification, some of the countries from both groups went to the previous World Cup. These effects appear to wear off fully in the year before qualification. In fact, the countries that went
to the previous World Cup were about 50% more aggressive in the years leading up to qualification than the countries that did not, although they were actually less aggressive prior to the previous World Cup.

The treatment effect seems to wear off after the second year following the World Cup, which is consistent with the qualitative evidence presented earlier suggesting that nationalism from sporting events can have downstream effects. However, this trend could partly be explained by the fact that the qualifiers started many conflicts in the first two years that reoccur in the third year. In fact, roughly 50% of the disputes started by the qualifiers between Year 2 and Year 3 were against countries that they attacked at least once since qualification. Put simply, when a group of countries experiences a large spike in aggression, it is likely to affect their aggression levels for several years, since some of the disputes that they start could lead to additional conflicts in the near future. Thus, part of the longevity of the effect is likely explained by this feedback mechanism.

The qualifiers not only took military action more often than the non-qualifiers, but the actions they took tended to be more violent. The Militarized Interstate Dispute dataset codes for the highest level of action taken by each country, with one being the threat to use force and 20 being the start of interstate war. In the two years after the World Cup, the median for the qualifiers was 15, whereas the median for the non-qualifiers was 11. Similarly, the qualifiers initiated seven disputes that resulted in fatalities, whereas the non-qualifiers initiated one. The Militarized Interstate Dispute dataset also codes for whether each dispute was intended to revise the status quo. In the two years following the World Cup, 72% of the disputes started by the qualifiers were revisionist, compared to only 54% for the non-qualifiers (p=0.001).

Table 4 shows the estimated treatment effects for different subsets of the data. The results are significant at the 1% level for the entire sample. They are also significant at the 1% level for the subset of countries where soccer is the most popular sport. On the other hand, there was no change in aggression for the subset of countries where soccer is not the most popular sport. This group includes the United States, Japan, New Zealand, Ireland, Australia, the United Arab Emirates, Jamaica, and Trinidad. Based on estimates from other studies, these results suggest that the World Cup increases state aggression about two-fifths as much as a revolution (Colgan 2010), and that it is comparable to electing a leader who is backed by the military (Lai and Slater 2006).

Table 4 also shows that the findings remain significant under various robustness checks, including tests that are insensitive to outliers and linear regression that controls for baseline differences between the two groups. The results are also robust to changes in the design. For instance, the findings remain significant when the two-point regression discontinuity window is set at one point or three points, as well as when ties are dropped. The results are also insensitive to shifting the five-point minimum score requirement and adjusting the time interval used for these tests. A full summary of these results and other robustness checks is available in the supporting information.
Table 4  Estimating the Effect of the World Cup on State Aggression

<table>
<thead>
<tr>
<th></th>
<th>Estimated Effect</th>
<th>p-value</th>
<th>n</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entire Sample</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.38**</td>
<td>0.006</td>
<td>142</td>
</tr>
<tr>
<td><strong>Sub-Groups</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Countries Where Soccer Is the Most Popular Sport</td>
<td>0.41**</td>
<td>0.010</td>
<td>132</td>
</tr>
<tr>
<td>Countries Where Soccer Is Not the Most Popular Sport</td>
<td>0.00</td>
<td>NA</td>
<td>10</td>
</tr>
<tr>
<td><strong>Shifting the Regression Discontinuity Window</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Countries That Qualified/Missed by One Point or Less</td>
<td>0.37*</td>
<td>0.023</td>
<td>92</td>
</tr>
<tr>
<td>Countries That Qualified/Missed by Three Points or Less</td>
<td>0.49**</td>
<td>0.002</td>
<td>162</td>
</tr>
<tr>
<td>Entire Sample (No Ties)</td>
<td>0.43*</td>
<td>0.010</td>
<td>102</td>
</tr>
<tr>
<td><strong>Other Statistical Tests</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Linear Regression with All Control Variables</td>
<td>0.39**</td>
<td>0.005</td>
<td>142</td>
</tr>
<tr>
<td>Permutation Test</td>
<td>0.38**</td>
<td>0.007</td>
<td>142</td>
</tr>
<tr>
<td>Standard t-test (not Difference-in-Differences)</td>
<td>0.44*</td>
<td>0.012</td>
<td>142</td>
</tr>
<tr>
<td><strong>Tests that are Insensitive to Outliers</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Signed-Rank Test</td>
<td>–</td>
<td>0.009</td>
<td>142</td>
</tr>
<tr>
<td>Dummy for Increase in Disputes Initiated</td>
<td>0.15**</td>
<td>0.004</td>
<td>142</td>
</tr>
<tr>
<td>Removing Great Powers</td>
<td>0.33*</td>
<td>0.018</td>
<td>139</td>
</tr>
<tr>
<td><strong>Other Outcomes</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Revisionist Disputes Initiated</td>
<td>0.38**</td>
<td>0.001</td>
<td>142</td>
</tr>
<tr>
<td>Disputes Initiated That Involved the Use of Force</td>
<td>0.28*</td>
<td>0.028</td>
<td>142</td>
</tr>
<tr>
<td>Disputes Initiated That Involved a Direct Attack</td>
<td>0.23*</td>
<td>0.020</td>
<td>142</td>
</tr>
</tbody>
</table>

Note: Unless otherwise specified, the results for these tests are from two-tailed difference-in-differences t-tests comparing the change in aggression between the qualifier and non-qualifier groups. I define change in aggression as the difference in the number of militarized interstate disputes initiated between (1) the period ranging from qualification to the second year after the World Cup and (2) the period of the same length prior to qualification. I use these time intervals to account for conflicts that may have been caused by the residual effects of nationalism. Nevertheless, the estimates presented here are similar for other choices of interval length. See the supporting information for a full summary of the robustness of these results.

* p<0.05, ** p<0.01, *** p<0.001
Figure 3 Change in Aggression for the World Cup

Note: The shaded regions represent the 95% confidence intervals, which were computed using non-parametric bootstrapping. The bandwidth for each graph was selected using the algorithm provided by Imbens and Kalyanaraman (2011), but the results remain significant for any $h \geq 1.5$.

Figure 3 shows the alternative way of conducting a regression discontinuity analysis. For each graph, the points on the right represent the means for the countries that qualified, and the points on the left represent the means for the countries that missed. The graphs show that the qualifiers experienced a significant increase in aggression in the two years following the World Cup. The difference at the cut-point is statistically significant for the middle graph and the one on the right.

Lastly, there is also some evidence that direct competition between countries at the World Cup increased the chances of conflict between them. The pairs of countries that played against each other found themselves on the opposite sides of military disputes about 60% more often in the two years following the World Cup compared to the two years before it ($n=758$, $p=0.020$). The number of these pairs that had at least one military dispute jumped from 32 to 47 (38.2% increase). On the other hand, the total number of all other disputes over these years increased by only 4.1%. These results suggest that the World Cup increased the likelihood of conflict between the pairs of countries that competed head-to-head.
As I discussed earlier, past historical evidence suggests that international sports lead to conflict because they incite mass nationalism. By doing so, they give leaders more incentive and opportunity to initiate military conflicts. Given the large estimated effect of the World Cup on state aggression found in the previous section, it is worth asking if there is also evidence that this mechanism is at work.

In fact, there are two noteworthy trends in the aggregate data that suggests that the World Cup led to more nationalism and militarism. Along with increasing state aggression, the World Cup also appears to have encouraged military participation and military spending. For military participation, the estimated effect is quite large. About 46% of the qualifiers experienced an increase in military personnel following the World Cup, whereas only 30% of the non-qualifiers did (p=0.047). Similarly, the countries that barely qualified experienced a much larger increase in military spending than the countries that barely missed (p=0.071). These results provide further evidence that the World Cup had an effect on mass nationalism as well as government decision-making.

There are also cases from the sample that suggest that World Cup nationalism is behind the results. The first is Senegal (2002), which initiated a dispute against Gambia about two weeks after its final World Cup game. In fact, the only time that Senegal has ever gone to the World Cup was in 2002, and the dispute that it started just afterward was the
only dispute that it was involved in from 1993-2010. This case is also striking because Senegal did surprisingly well at the World Cup. In the group stage, it defeated France, the defending World Cup champion and its former colonizer. The victory was so remarkable that CNN lists it as one of the ten greatest upsets in World Cup history. Senegal then beat Sweden in the first round of the knockout stage, making it just the second African team to advance to the quarterfinals. It lost its next game to Turkey 0-1, exiting the tournament on June 22, 2002.

No doubt, the World Cup created a powerful surge of nationalism in Senegal, which began immediately after qualification. Senegal’s president, Abdoulaye Wade, was visiting Jacques Chirac in France at the time, but returned home early to celebrate Senegal’s qualification for the World Cup. As he explained, “it’s the most important thing that can happen to any country and I will join the team and the nation in celebrating by reducing the amount of time I was expected to stay in Paris” (Ralph 2007). To reward his players, he invited them to his palace and presented each of them with bonuses of about $15,000. Nation-wide celebrations also broke out after Senegal defeated France and Sweden, and they were encouraged by the government. For instance, after the win over France, President Wade declared a national holiday and paraded around the capital in a vehicle with the top open so that fans could see him juggling a soccer ball.

On July 7, 2002, just 15 days after losing to Turkey in the knockout stage, Senegal initiated a military dispute against Gambia. Specifically, it asserted its control over the Casamance region in southern Senegal by expelling a group of rebels who were allegedly supported by Gambia. Senegal then deployed troops along the Gambian border to ensure that the rebels would not reenter the country. This example featured clear military action, but it is most notable because of its timing. The fact that it is the only dispute that Senegal initiated in an almost twenty-year period and that it happened just after Senegal made its only World Cup appearance in history suggests that sports nationalism can encourage leaders to make aggressive foreign policy moves.

A second notable case from the sample is Cameroon (1994), which illustrates how the World Cup can lead to reoccurring disputes. In October of 1993, Cameroon qualified for the World Cup, causing celebrations to erupt throughout the country. In the following March, about three months before the World Cup was set to begin, Cameroon invaded the Bakassi Peninsula, an oil-rich region located between Cameroon and Nigeria. This dispute was the first that Cameroon was involved in since 1987. It resulted in a series of clashes between Cameroon and Nigerian troops in the following years, two of which were coded as new disputes initiated by Cameroon (December 1995 and March 1996). This case demonstrates how one dispute can lead to a series of conflicts, which may help explain the longevity of the effect.

These cases also suggest another possible mechanism that may be at work. In some cases, leaders may not just be reacting to changing public opinion; they may be passionate fans themselves who become more nationalistic from watching their teams play. No doubt,
it is common for government leaders to attend games, and some even schedule their cabinet meetings around them so that they can watch their teams play on television (Markovits and Rensmann 2010). After all, most state leaders are men, who are much more likely to be sports fans than women. Simply watching games could therefore encourage some leaders to see other states as competitors, which would bypass the public opinion mechanism.

Of course, it is very difficult to assess this mechanism empirically, since it requires knowing leaders’ inner thoughts. For instance, Senegal’s president seemed passionate about soccer, but it all might have been for show. There are also clear examples of leaders like Hitler who did not care about sports for their own sake, and who were just interested in how they could be used to manipulate public opinion. Thus, this mechanism cannot explain all of the conflicts that have occurred because of sports nationalism, but it may explain some of them, and it is therefore worth taking seriously.

**CONCLUSION**

This study provides strong evidence that surges of nationalism can have a substantial impact on interstate conflict. In doing so, it confirms an important causal relationship that has been assumed in much international relations scholarship, but had yet to be empirically established. Put simply, the history of international sports provides enough qualitative and quantitative evidence to conclude that fluctuations in nationalism can have a large impact on state aggression.

In short, nationalism should continue to play an important role in the study of international relations. The premise that nationalism causes conflict can be very useful for interpreting history, constructing theories, and making arguments to policymakers. More work is also needed to identify other potentially dangerous sources of nationalism. One promising avenue would be to look at whether major national achievements and tragedies tend to increase state aggression in the short run. Some research has already been done on how nationalism from September 11 made many Americans more willing to invade Iraq (McCartney 2004), but there is still room for a broader theory that takes into account other events that cause temporary spikes in nationalism.

This study also has some important policy implications. Since international sports are a powerful source of nationalism, we should try to limit the ability of certain leaders to use that nationalism to increase domestic support for their aggressive foreign policies. For example, allowing Putin to host the 2014 Winter Olympics and 2018 World Cup was a poor decision, and international sports organizations should be careful not to make similar mistakes in the future.

We might also consider having some sporting events where countries play as small regional blocks like a Scandinavian team and Balkans team. This format would encourage people to identify with their regions rather than their nations, and could therefore improve how people perceive neighboring states. They would be allies rather than competitors. Moreover, this format would ensure that a country’s opponents would typically be groups
of countries that were further away geographically and less likely to be possible targets of future military aggression. Thus, this format would probably have a much more positive effect on international relations than the current practice of constantly pitting countries against each other on the international stage.
REFERENCES


Hitchens, Christopher. 2010. “Why the Olympics and Other Sports Cause Conflict.” *Newsweek*. 

27


United We Fight: Democratic Unity and State Aggression

Andrew Bertoli

ABSTRACT. Are democratic leaders more likely to use force abroad when their co-partisans control the main bodies of government? Past studies have sought to answer this question with regression analysis, but their statistical models lead to incompatible results. I shed new light on this debate using a regression discontinuity design. Looking at democracies from 1815-2010, I compare cases where political parties barely gained control of government to cases where they barely fell short of doing so. The results provide compelling evidence that government unity increases state aggression. Specifically, I find that united government makes countries much more likely to initiate high-level disputes. The effect is particularly notable for the United States. These results underscore the importance of being alert to the impact of domestic political developments on foreign policy.

In August of 2013, Republican leaders in the House of Representatives announced that Obama’s request to use force against Syria would not pass. Obama had asked the Republican-controlled House for authorization to retaliate against Assad for his use of chemical weapons in the Syrian Civil War. While it is not necessary for presidents to receive permission from the House for conflicts lasting shorter than 60 days, it would have legitimized Obama’s decision, ensured that Republicans would not try to punish him for acting without their approval, and allowed him to stay involved in the conflict if it lasted longer than two months. However, the normally hawkish Republicans refused to authorize the military strike. As Republican Representative Tim Griffin explained, “I am reluctant to give him a license for war when, with all due respect, I have little confidence he knows what he is doing.”

Clearly, circumstances were very different from the situation in 2009 and 2010, when the legislative branch was controlled by the Democrats. During this period, Obama had a substantial amount of freedom when it came to foreign policy, allowing him to shift military troops from Iraq to Afghanistan and expand the drone program without having to worry about congressional resistance.

This case raises a key question about the impact of domestic factors on foreign policy: Are democracies more likely to take military action when their governments are united,

I would like to thank Jasjeet Sekhon, Ron Hassner, Thad Dunning, Aila Matanock, Michaela Mattes, Fredrick Salve, Josh Kalla, Caroline Brandt, Nina Kelsey, Jason Klocek, Anne Meng, Tara Buss, Alice Ciciora, Elizabeth Herman, Melissa Carlson, Bora Park, and Jay Varellas for helpful comments.

meaning that their executive and legislative branches are both controlled by a single party? No doubt, there are compelling reasons to think that united government may make countries more aggressive. On one hand, leaders of united governments may be less constrained when it comes to foreign policy, making it easier to take military action against other states (Howell and Pevehouse 2005). Additionally, they may feel that they are less likely to be punished by the legislature for taking unpopular military action. However, there are also reasons to think that united government might not make aggression more likely. National security information is often kept secret, so decisions by leaders to use force may be largely uninfluenced by the legislature (Rosato 2003). Leaders facing strong opposition parties might also be unable to advance their agendas at home, giving them more reason to focus on international relations where they may have more autonomy (Brulé 2006).

This question has important implications for our understanding of international relations. First, one proposed explanation for the democratic peace is that strong opposition parties make it difficult for leaders to take military action against other countries. If democracies tend to behave more aggressively when they are united, then it would provide evidence for this important causal mechanism while also pointing to a caveat of the theory: that the democratic peace is more likely to hold when democracies are divided. Second, if government unity increases state aggression, then policymakers and the public should be alert to the potential danger of unnecessary conflicts when one party is in power.

But despite the importance of this question, past researchers have been unable to reach a consensus on how united government affects state aggression. Their studies, which rely primarily on regression analysis, come to a wide range of competing conclusions. Many find that united government increases state aggression (Clark 2000; Clark and Nordstrom 2006; Brulé 2006; Pevehouse 2005). However, others find little evidence that this is the case, instead concluding that politics stop at the water’s edge (Gowa 1998; Schultz 1999; Leblang and Chan 2003). Put simply, efforts to resolve this question with statistical models have led to incompatible results.

I shed new light on this debate by taking advantage of a natural experiment where government unity was assigned to countries as-if randomly. Specifically, I compare cases where political parties barely gained control of their governments to cases where they barely fell short of doing so. The idea is that when parties are on the verge of achieving united government, it is basically a coin-flip which ones barely succeed and which ones barely fail. This conjecture is likely true because large national elections have many unpredictable elements. Thus, it would be very difficult for parties to know exactly how close they were to united government beforehand and precisely manipulate turnout to get just enough votes to take control of government. In short, when countries are at the threshold, whether they become united or divided is largely left up to chance.

The results provide compelling evidence that united government increases state aggression. While the countries with barely united and barely divided governments were very similar prior to the elections, the countries with barely united governments became much
more aggressive after their governments came to power. Specifically, they were much more likely to initiate high-level military disputes against other states. However, they were not significantly more likely to initiate low-level disputes.

This paper proceeds as follows. Section 1 considers several theoretical arguments about the relationship between united government and state aggression. Section 2 reviews the observational evidence. Section 3 lays out the research design in more detail and verifies that the countries with barely united and barely divided governments were very similar on pre-treatment characteristics. Section 4 analyzes the results with a number of statistical methods. Section 5 considers the broader implications of these findings. The final section concludes.

**SECTION 1: THEORY**

The argument that united government increases state aggression rests on three basic ideas. The first is that opposition parties will constrain leaders when government is divided, specifically because they have the opportunity and incentive to do so. The second is that leaders will not become less interested in foreign policy during times of united government, say because they will be much more focused on passing domestic legislation through a favorable legislature. The third is that the countries that are the potential targets of aggression will not be more willing to make concessions when facing a united government, because they realize that they are negotiating with a less constrained adversary. Assuming that these conditions hold, united government should tend to increase the likelihood of state aggression. However, if even one of these conditions is typically untrue, then the effect of united government could be negligible.

**Opposition Parties.** Let us start with the question of whether opposition parties can better constrain leaders when government is divided. The most obvious way that they could do so is by refusing to give leaders permission to use military force abroad (Brulé 2006). For instance, if the legislature is responsible for approving any military engagement, then opposition parties in divided governments may simply be able to vote down such policies when leaders request them. Thus, the legislature could act as a veto player that simply denies the leader permission to use force (Tsebelis 1995).

However, it is typically not this easy for opposition parties to constrain leaders. Although most democratic legislatures are formally responsible for authorizing military action, leaders often launch military strikes without legislative approval. For instance, the United States government used force against other countries in 126 separate military disputes from 1815-2010, but only 13 of these cases were actually authorized by Congress, and only four involved a formal declaration of war (French and Bradshaw 2014). This ability of leaders to bypass the legislature is partly explained by the need to make decisions quickly in international crises and keep vital national security information secret (Rosato
2003). Thus, the formal institutional power that legislatures are usually granted over the decision to use force often fails to keep executives in check.

A more likely way that the opposition can exert influence in divided government is by threatening to punish leaders. Strong opposition parties can often defund unwanted military campaigns, remove leaders from office, or refuse to cooperate in other policy areas. Thus, leaders who face a divided government might realize that acting against the wishes of the opposition would be costly and opt not to use military force. This type of deterrence would primarily apply to cases that involve the actual use of force against other states. After all, the legislature would not be able to defund or credibly punish very low-level disputes that merely involve saber-rattling. Thus, we should expect that a strong opposition party could, at most, prevent leaders from taking significant military action against other countries.

Of course, even if opposition parties have the power to constrain aggression, they must also have incentives to do so. Whether they do will largely depend on how much their foreign policy goals differ from those of the leader. For instance, if one party had much more dovish foreign policy preferences than the other, then we might expect military conflict to occur more when the hawkish party achieves a united government. Another case of competing foreign policies could arise if both parties were equally willing to use force but had very different interests. They might favor different allies or fear different foreign enemies, and thus block each other’s military initiatives whenever possible.

However, even if the leader and the opposition agree on when force should be used, opposition parties might have other reasons to prevent leaders from taking military action. For example, they might doubt that the leader is competent enough to manage the conflict effectively, as Representative Tim Griffin claimed when Obama requested authorization to attack Syria. Another possibility is that the opposition might fear that foreign conflict would improve the leader’s approval rating. A number of scholars have shown that interstate conflict often produces a “Rally ’Round the Flag Effect” that increases the executive’s popularity (Mueller 1970; Oneal and Bryan 1995). By doing so, it may strengthen the leader’s party at the expense of the opposition, since popular executives can improve the electoral prospects of their co-partisans in the legislature (Campbell and Summers 1990; Howell and Pevehouse 2005). Thus, even when the members of the opposition largely share the leaders’ foreign policy goals, they may still fear that military conflict will ultimately hurt them in the polls.

Leaders. A second condition that must hold is that united government cannot distract leaders from foreign policy. If it encourages them to turn their attention to domestic politics, where they will have an easier time passing legislation, then united government might make countries less aggressive.

However, this logic assumes that leaders often struggle to focus on multiple things at once, even when the conditions make advancing their agenda relatively easy. Given that
they have the manpower and resources to research many issues simultaneously, it is unlikely that greater opportunity domestically would divert attention away from foreign policy. In fact, leaders might not need to spend as much time on domestic policy when government is united, because they will not have to negotiate with the other party on every piece of legislation. Thus, there is little reason to think that united government will cause leaders to pay less attention to foreign policy.

Target States. Lastly, because military conflict often results from one country refusing to give into another country’s demands, it is also important to consider how united government might influence the strategy of a democracy’s adversaries. If these states recognize that they are negotiating with a less constrained leader, they might be more willing to make concessions that prevent conflict from breaking out.

However, there are several reasons for believing that target states would not be more conciliatory when facing a united government. First, the target country might not realize that the democratic leader is in a better position to use force against it. Given that political scientists disagree about whether united government gives leaders more freedom in foreign policy, it is likely that many target countries are also unsure whether they are negotiating with a less constrained state. Second, even if the leaders in the target country understood the increased possibility of being attack, they might still prefer to risk conflict than make concessions. For instance, they might fear how looking weak on the international stage would hurt them in domestic politics, or they might be very confident in their chances of winning a fight. Third, the target country might doubt the resolve of the united democracy if the opposition parties in that democracy, unable to prevent the conflict through legislative action, voice their opposition to the conflict in the media. Thus, there is good reason to think that target states will not necessarily be more willing to make concessions when facing a united government.

**SECTION 2: EXISTING EVIDENCE**

Existing evidence about the relationship between united government and state aggression is unclear. On one hand, the U.S. record of major war suggests that government unity has a large impact on a country’s likelihood of engaging in large-scale military conflict. Since the Revolutionary War, there have been nine interstate wars that involved at least 1,000 U.S. battle deaths. The American government was united before entering all but one of them—the recent war in Afghanistan following September 11. As Figure 1 shows, this number is strikingly large given that the United States has only been united for about half of its history. In fact, the wars started under united government resulted in a total of 434,609 American battle deaths, while the war started under divided government resulted in 1,724 American battle deaths.
However, this correlation between united government and international conflict might be explained by confounding factors rather than a true causal relationship. For example, publics might be more willing to elect a united government when the international system seems particularly dangerous. To overcome this problem, a number of studies have used regression analysis to control for possible confounders. However, these studies come to competing conclusions. For example, Clark and Nordstrom (2006) find that united government increases the likelihood that countries will initiate military disputes, while Schultz (1999) finds little statistical evidence that opposition parties effectively constrain leaders from initiating military disputes. Leblang and Chan (2003) reach similar conclusions when investigating involvement in interstate wars.

Even research that just focuses on the United States has failed to reach agreement. Gowa (1998) finds that united government has a negative and insignificant effect on U.S. military disputes. On the other hand, Clark (2000) and Brulé (2006) conclude that government unity increases the likelihood that the United States will initiate military disputes. Meanwhile, Howell and Pevehouse (2005) find that the United States tends to initiate more high-level disputes under united government, but that this pattern does not hold for low-level disputes. Put simply, regression analysis has failed to produce a clear answer, which is not entirely surprising given that regression results can vary widely based on model specification and the variables that researchers choose to control.

The ideal way to get around this problem would be to run an experiment where we took a large group of countries and randomly assigned some to be united and others to be divided. Randomization would get around the confounding problem because it would create balance across observable and unobservable characteristics. Thus, it would allow for us to make a fair comparison between countries with united and divided governments. If the group of countries with united governments behaved much more aggressively, we could conclude that this difference was due to government unity rather than some confounding factor.

Of course, we cannot run a real experiment to answer this question, because assigning a large number of countries to be united or divided is infeasible. However, the process
of national elections within democracies created a natural experiment that is very similar to the ideal study described above. I turn my attention to this natural experiment in the following sections.

SECTION 3: RESEARCH DESIGN

Strategy for Identifying the Causal Effect. This paper tests how united government affects state aggression by using a regression discontinuity (RD) design. RD is a quasi-experimental approach that can be used when a treatment is given to units that surpass an important cut-point in a scoring system (Dunning 2012). The idea is to compare the group of units that scored just above the cut-point to the group that scored just below it. As long as there is some randomness in the scoring process, it should be close to random which units end up on either side of the cut-point (Lee 2008). Thus, the units just above the cut-point should be, on average, similar to the units just below the cut-point, except that the units above received the treatment and the units below did not.

In political science, RD is most often used to estimate the effects of winning close elections on a variety of outcomes, including personal wealth and the probability that a person or party will win the next election (Eggers and Hainmueller 2009; Titiunik 2009; Broockman 2009; Gerber, Kessler, and Meredith 2011; Eggers, Folke, Fowler, Hainmueller, Hall, and Snyder 2013). However, RD has also been used to study questions in international relations, like how being elected to the UN Security Council affects the willingness of countries to contribute to peacekeeping missions (Voeten 2013), how barely qualifying for the World Cup affects state aggression (Bertoli 2016), and how electing a younger leader affects a country’s likelihood of engaging in military conflict (Bertoli, Dafoe, and Trager 2016).

For this study, I compare cases where parties barely gained control of government to cases where they barely fell short of doing so. For instance, in 1990, the strongest party in Costa Rica, the Social Christian Unity Party, won the presidency by about 4% of the vote and gained control of the parliament by exactly one seat. However, when the next election came around in 1994, the situation was reversed. The strongest party (this time the National Liberation Party) won the presidency by 2% of the vote but fell short of gaining control of the parliament by exactly one seat. In very close national elections like these, whether countries ended up barely united or barely divided should be essentially random. Thus, the group of countries that where barely united should be, on average, pretty similar to the group of countries that were barely divided, except for the key difference in government unity.

Democracy and Election Data Sources. To construct the dataset, I began by identifying every democracy in the international system from 1815-2010. Following past studies (Schrock-Jacobson 2012; Marshall, Gurr, and Jaggers 2013), I counted a country as
a democracy in a given year if it had a POLITY IV Institutionalized Democracy Score above five. After creating the list of democracies, I collected election data from a number of sources. The key volumes were *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). I also had to use some online databases to get information on elections in recent years. These sources included the International Foundation for Electoral Systems’ Election Guide, the African Election Database, and the European Election Database.

**Scenarios.** There are three basic ways that power is shared between different government bodies in a democracy. First, some countries have an independently elected executive branch and exactly one legislative body (unicameral presidential systems). These countries are united if one party controls the executive and legislative body, and they are divided otherwise. Countries with this form of government include Austria, France, Finland, Ireland, and South Korea. Second, some democracies have two legislative bodies, such as a house and senate, while lacking an independently elected executive branch (bicameral parliamentary systems). In these countries, the executive is typically elected by one of the two legislative bodies. Countries with this government structure include Australia, Belgium, Italy, Japan, and Spain. These countries are united if one party controls both legislative bodies. Third, some countries have an independently elected executive branch and two legislative bodies (bicameral presidential systems), like the United States, Brazil, Mexico, Colombia, and the Philippines. In order for these countries to be united, a party must control all three of these government organizations. I exclude from this analysis democracies that have only one legislative branch that chooses a prime minister (unicameral parliamentary systems), since these countries are always united under whatever party or coalition controls the legislature.

It is fairly straightforward to think about how close governments are to being united in the three main contexts described above. Figure 2 illustrates government control for countries with a president and one legislative body as a two-dimensional RD. Each point represents the level of control for the strongest party in the country. Points to the right of the vertical axis are cases where the party controlled the legislature, and points above the horizontal axis are cases where the party controlled the presidency. Thus, the upper right-hand quadrant contains the cases where the most powerful party controlled both the executive and legislative bodies (united government), and all other quadrants contain divided governments. Countries that were barely united or barely divided are the ones that are very close to being inside or outside of the upper right-hand quadrant. Similarly, Figure 3 depicts countries with two legislative bodies and no independently elected president. Again, points in the upper right hand quadrant represent united government, while points in all other quadrants depict divided governments. Countries with an independently elected executive and two legislative bodies must be thought of as a three-dimensional
Figure 2. Strength of Most Powerful Party in Elections with a President and Legislature

Figure 3. Strength of Most Powerful Party in Elections with Two Legislative Bodies
regression discontinuity (Figure 4), where governments are united when a party surpasses all three cut-points.

**Countries.** Figure 5 shows the countries in the dataset. I include all democracies from 1815-2010 that have at least two elected branches or chambers and were close to being united or divided at least once. Most of these countries have a president and legislature with either one or two chambers, although there are also some countries that have just a bicameral legislature. I did not include ceremonial legislatures that have little or no formal authority, such as the German Bundesrat and the United Kingdom’s House of Lords. To determine whether a legislative body is ceremonial or has formal power, I referred to *The Handbook of National Legislatures*, by Fish and Kroenig (2009).

A potential concern here is that some of these democracies were not actually democratic, such as the United States in the nineteenth century. Slavery was legal in the South until the end of the Civil War, and women did not have the right to vote until 1919. While this
point is legitimate, the key reason that I focus on democracies is their fair elections, not their human rights records or who they allow to vote. In fair elections, political parties should not be able to manipulate their vote share to barely achieve united government, so we have strong reasons to believe that united government was essentially random in close cases. On the other hand, in non-democracies, political parties may be able to rig the election in their favor. In fact, there is good evidence that parties were able to manipulate their vote shares when they were close to united government, which I present in the Online Appendix. In contrast, parties in democracies appear to have been unable to do so, as I will show in the following sections.

**Forcing Variable.** Regression discontinuity analysis requires a well-defined forcing variable, or measure for how close units are to the cut-point. I define the forcing variable following Bernecker’s (2014) working paper about the effect of united government on welfare reform in U.S. states between 1978 and 2010. To my knowledge, his paper is the only other that attempts to uncover a natural experiment about united government using regression discontinuity.
His way of defining the forcing variable is intuitively appealing. For united governments, the value of the forcing variable is simply the shortest linear distance to the cut-point when the election is represented graphically. Thus, if Republicans in the United States gained control of the Presidency by 3%, the House by 6%, and the Senate by 7%, the value of the forcing variable would be 3%. For divided governments, the value of the forcing variable is the sum of the linear distances from the cut-point in all government bodies where the strongest party lacks control. For instance, if the Republicans won the Presidency by 9%, but lost the House by 2% and Senate by 1%, then the value of the forcing variable would be \(-2\%+(-1\%)=-3\%\). This metric is called the \(L_1\) norm (or taxicab distance), and it is formally written as

\[
d_1(a, b) = \sum_{i=1}^{n} |a_i - b_i|
\]

where \(a\) is the coordinates of the country-year and \(b\) is the coordinates of the closest point to \(a\) on the cut-point.

In cases where one or more government bodies were not up for a vote, such as midterm elections in the United States, no more than one party can be in a position to achieve united government. For example, if the Democrats controlled the Presidency, the Republicans would have no way of getting a united government in the midterm elections. In these cases, I use the minimum distance to the cut-point for the only party that can achieve a united government. Thus, if the Democrats lost the House and Senate by 6% each in the midterm, then the value of the forcing variable would be \(-12\%\). Cases where no party could achieve a united government were dropped from the analysis.

Figure 6 plots this forcing variable for the countries in the sample. There is no evidence that countries tended to sort toward being barely united or barely divided, which is what we should expect if treatment assignment is essentially random around the cut-point. In fact, the distribution is very smooth through the cut-point, suggesting that countries did not sort. This graph also shows that the sample size is not small. Overall, there are 257 observations within 10% of the cut-point, and 66 observations within 2% of the cut-point. Thus, we have enough data to estimate a treatment effect with some precision.

**Other Methodological Issues.** Before moving on, there are two final issues that need to be addressed. First, there were 7 cases where a party won exactly 50% of a legislative body, raising the question of whether this counted as control. Fortunately, in three of these cases there were clear tie-breaking rules that gave the advantage to the party that controlled the executive branch. I count the four remaining cases as instances where countries fell somewhere between united and divided government, and therefore leave them out of the analysis. However, the main results are robust to including them in either the treatment or control group. Second, there were some cases where governments were barely united or barely divided following elections, but they changed afterward because some politicians switched parties. Fortunately, these cases are exceptionally rare. However, reclassifying
them could result in bias, since then certain countries could sort to united or divided post-election. Therefore, a country’s treatment status must be based solely on the election results, which can be thought of as a very strong instrument for government unity.

For presidential elections in the United States, I used votes in the electoral college rather than popular vote, since candidates are required to get a majority of votes in the electoral college. This method might result in a decisive election being coded as a close one if there was a case where each candidate won every state by a large margin, and one candidate was only slightly ahead of the other in the electoral vote. However, there are no cases like that in American history. However, in the three elections when the U.S. presidential candidates were within 5% of each other in the electoral college (1876, 1916, and 2000), there were multiple swing states that could have caused the winner to lose the election had the results been slightly different. Therefore, using the electoral college does not lead to any decisive elections to be coded as close.

**Estimation.** To estimate the effect of united government on state aggression, I take the standard approach when dealing with continuous forcing variables. Specifically, I use a local linear smoother to estimate two regression lines on either side of the cut-point and then look at the difference between these two lines at the cut-point. The resulting estimator can be thought of as the difference between the average level of aggression for countries
with united governments and the average level of aggression for countries with divided
governments, as it is defined exactly at the cut-point. To carry out this estimation, I use
the “rdrobust” software developed by Calonico, Cattaneo, and Titiunik (2014), which pro-
duces bias-corrected estimates with robust standard errors. Their software also calculates
the optimal bandwidth to minimize the mean-squared error, which is the standard crite-
rion for bandwidth selection. To illustrate the treatment effect, I also create the normal
regression discontinuity graphs with bootstrapped standard errors. These graphs do not
show the bias correction, but the results are very similar to the bias-corrected estimates.
To make this estimator doubly robust, I use the change in aggression for each country
(difference-in-differences) rather than just its unadjusted aggression level.

**Measuring Aggression.** Similar to past studies (Gartzke 2007; Dafoe, Oneal, and Rus-
sett 2013), I measure aggression using the number of militarized interstate disputes (MIDs)
that a country initiates. These disputes are instances where states explicitly threaten, display,
or use force against other countries (Ghosn, Palmer, Bremer 2004). Furthermore, I
distinguish between high-level disputes that involve the use of force and low-level disputes
that do not. This distinction is important because there are theoretical reasons to expect
united government to have a larger effect on high-level disputes than low-level disputes.
Since the amount of time between elections varies across cases, I also divide the number of
disputes in each case by the total number of years that the government remained in power.
Thus, the unit of measurement is disputes-per-year for each country.

**Checking for Balance.** Before moving on to the main results, we can check whether the
estimator identifies systematic imbalances for any important covariate. Figure 7 shows the
balance tests for a number of key pre-treatment characteristics. Overall, the balance looks
better than what we would expect in a typical randomized experiment, where p-values
would be normally distributed between 0 and 1. The only variable that is close to being
imbalanced is low-level disputes, primarily because the bare non-qualifiers initiated sur-
prisingly few low-level disputes in the previous period. The two groups were comparable
on high-level disputes, which is the main outcome of interest in this study. Moreover,
when high and low-level disputes are combined, the two groups are fairly well balanced
(p=0.24).

**SECTION 4: FINDINGS**

Figure 8 tracks the aggression levels of the countries that were within 2% of the cut-
point. This sample consists of 33 countries that were barely united and 33 that were barely
divided. The graph shows that the barely united countries initiated many more disputes
than the barely divided ones, even though the two groups had similar aggression levels in
the prior period. On average, the barely united countries initiated about 0.38 disputes per
Figure 7. Testing for Balance at the Cut-point

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Treatment Mean</th>
<th>Control Mean</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\ln(\text{Iron and Steel Production}))</td>
<td>9.07</td>
<td>8.95</td>
</tr>
<tr>
<td>(\ln(\text{Military Expenditures}))</td>
<td>13.6</td>
<td>13.3</td>
</tr>
<tr>
<td>(\ln(\text{Military Personnel}))</td>
<td>4.53</td>
<td>4.53</td>
</tr>
<tr>
<td>(\ln(\text{Energy Consumption}))</td>
<td>10.8</td>
<td>11.4</td>
</tr>
<tr>
<td>(\ln(\text{Total Population}))</td>
<td>9.95</td>
<td>10.4</td>
</tr>
<tr>
<td>(\ln(\text{Urban Population}))</td>
<td>8.76</td>
<td>9.1</td>
</tr>
<tr>
<td>Previously United</td>
<td>0.642</td>
<td>0.679</td>
</tr>
<tr>
<td>Previous Low-Level Disputes</td>
<td>0.367</td>
<td>0.02</td>
</tr>
<tr>
<td>Previous High-Level Disputes</td>
<td>0.168</td>
<td>0.118</td>
</tr>
</tbody>
</table>

\(p\)-value

year after their governments came to power, whereas the barely divided countries initiated roughly 0.19.

Figure 9 presents the bias corrected estimates with robust confidence intervals, based on Calonico, Cattaneo, and Titiunik (2014). These tests are the same as the ones used to check for balance at the cut-point in Figure 7, which found no significant differences for any pre-treatment covariate. The outcomes are the change in disputes between the period when the governments were in power and the period before. While there appears to be an increase in the total number of disputes initiated, this result is not significant at the 5% level (p=0.44). However, differentiating between high-level and low-level disputes provides much more conclusive results. Specifically, the estimates indicate that united government makes countries much more likely to initiate high-level disputes (p=0.02), while having no discernible impact on low-level disputes (p=0.57). Substantively, the estimates suggest that changing from divided to united government causes countries to initiate about 0.38 more high-level disputes per year.
Figure 8. Change in Aggression for Barely United and Barely Divided Countries

Note: This graph shows the disputes initiated by countries that were within 2% of being united or divided.

**Individual Countries.** To illustrate which countries are most responsible for influencing the results, I have provided a standard regression discontinuity graph (Figure 10) where the observations close to the cut-point are labeled. Specifically, I labelled any country that had at least one high level military dispute and fell within the bandwidth ($h = 1.83\%$). This bandwidth was selected using the algorithm provided by Calonico, Cattaneo, and Titiunik (2014). The graph shows that the United States is playing an important role in influencing the results, but no other country stands out. In fact, the results are almost significant for just the United States ($p=0.053$).
This figure also shows that the barely divided governments tended to experience a decrease in aggression. While this drop might seem surprising, recall that many of these countries had united governments in the previous period. Therefore, since they went from united to divided government, we should expect their aggression levels to drop if united government increases state aggression. It also makes sense that countries further away from the cut-point experienced little change in aggression. These countries tended to have the same treatment status in the previous period. That is, they tended to go from united to united or divided to divided, so we should not expect their aggression levels to change much. Only the countries that were close to the cut-point had a high probability of switching from united to divided or divided to united, so that is where we should expect to see the largest change in aggression levels.

**Robustness Checks.** The results for high-level disputes also remain significant for a number of important robustness checks. For instance, they hold after controlling for the pre-treatment characteristics in Figure 7 (p=0.047). Adjusting for these factors can be done by predicting the aggression levels for each unit using a linear model and then plugging the residuals from this model into the regression discontinuity tests in place of the outcomes. This procedure is commonly used as a robustness check for regression discontinuity designs, as it verifies that baseline differences between the two groups are not explaining the results. The results are also significant for a simple t-test that compares the countries that were within 0.5% of the cut-point (p=0.033). Moreover, these countries are almost perfectly balanced on high-level disputes in the previous period (p=0.960).
Figure 10. Identifying the Countries That Are Influencing the Results

Note: The bandwidth for this graph was selected using the algorithm provided by Calonico, Cattaneo, and Titiunik (2014). The shaded regions represent the 95% confidence intervals, which were computed using non-parametric bootstrapping.

In fact, the broader historical records of aggression between the barely united and barely divided countries were very similar, except for the periods when the newly elected governments came to power. Figure 11 shows the distribution of the regression discontinuity test statistic going 40 periods back and 10 periods forward. The only time that the barely united countries behaved significantly more aggressively than the barely divided countries was immediately following the election that left them barely united or barely divided. This balance demonstrates that the barely united countries did not tend to be more aggressive countries historically. The treatment effect appears only right after the countries received the treatment.

47
This study finds strong evidence that united government increases the likelihood of interstate conflict. While the results are fairly straightforward to interpret, there are several important points that are worth keeping in mind about this finding.

First, it does not imply that increasing government unity will tend to make state aggression more likely for all types of countries. After all, the countries considered in this study are all democracies. Whether an exogenous increase in government unity would have a similar effect on an autocratic regime is a different question. In fact, the literature on diversionary war theory suggests that government unity may actually decrease aggression for non-democracies, since it alleviates leaders’ need to use international rivalry to strengthen their support at home (Oakes 2006). Another group of countries that may respond differently are democracies that are experiencing severe economic downturns, as their leaders may find military action an easier way to win public support than trying to pass legislation through a divided government (Brulé 2006). Thus, it seems very unlikely that government unity increases state aggression under all conditions. The results of this study simply indicate that democracies typically tend to resort to military force more often when their governments are barely united.

Second, these findings suggest that opposition parties play an important role in constraining democratic aggression. In fact, we might expect that democratic peace theory
will be more likely to hold for democracies with divided governments. Of course, other factors may also contribute to the democratic peace, such as liberal norms, open international markets, and better information about government capabilities and intentions. However, the role of strong opposition parties in constraining democratic aggression should not be overlooked. As this study shows, foreign policy is more constrained when opposition parties possess greater institutional power.

Third, these results suggest that the political opposition often opposes the military ambitions of leaders. This suggests that opposition parties may be concerned about the “Rally ’Round the Flag” effect that could strengthen the power of the leader’s party at their expense. Thus, the results might be best understood not simply as a one-way effect of domestic politics on international relations. Rather, they may involve the interaction between these two realms. Events at the international level can affect outcomes at the domestic level, which means that domestic actors could have incentives to shape those events and thereby influence international relations.

**CONCLUSION**

While united government may lead to positive outcomes at the domestic level, scholars have long suspected that it has adverse effects on international relations. Past research on this topic provided compelling reasons for why institutional unity might matter, but different statistical models led to competing conclusions about the true effect of united government on interstate conflict. This study shed new light on the debate by taking advantage of a natural experiment where united government was assigned to countries as-if randomly. The results provide strong evidence that united government increases the likelihood of state aggression, especially when it comes to high level-disputes.

This study should make policymakers and the general public much more wary about the potential for unnecessary conflicts when government is united. At the same time, it should draw attention to the possibility that divided governments might avoid some necessary conflicts, since opposition parties may wrongly oppose aggressive foreign policy moves that would actually be in the best interests of their countries. This danger seems to be especially concerning for the United States, which influences the results more than any other country. In sum, Americans should be acutely aware of the potential for domestic politics to have a negative impact on foreign policy.
REFERENCES


French, Bill and John Bradshaw. “Ending the Endless War An Incremental Approach to Repealing the 2001 AUMF.” Published by the National Security Network.


Regression Discontinuity in Practice: Solving Common Problems in Applied Research

Andrew Bertoli

ABSTRACT. Although regression discontinuity (RD) is a very simple research design in theory, many applications are riddled with subtle complexities, even when treatment assignment is random close to the cut-point. In this paper, I examine three of the most common problems for regression discontinuity designs: (1) variation in competitiveness across strata, (2) units that appear in the sample multiple times, and (3) independent variables of interest that are not actually assigned by the RD. In each case, I explain the problem formally, offer straightforward solutions, and illustrate the main points with simulations and real-world data. To increase clarity, I treat regression discontinuity as a natural experiment in a narrow window around the cut-point. While this approach differs from the more standard practice of taking units’ scores as fixed and estimating the difference at the cut-point using two regression lines, I show that the two approaches are very similar mathematically.

Over the last decade, regression discontinuity (RD) has gained popularity in many scientific fields. Not only is it one of the most trusted tools of causal inference short of an experiment (Bernardi and Skoufias 2004; Green et al., 2009), but its simplicity allows researchers to present their findings to other scholars and the general public much more effectively (Hopkins and McCabe 2012; Hall 2015). Given its virtues, the popularity of the regression discontinuity design is likely to continue to grow for the foreseeable future, especially considering the prevalence of scoring systems with cut-points. In recent years, it has even spread to fields like sociology (Rao, Yue, and Ingram 2011; Legewie 2013; Bernardi 2014), international relations (Voeten 2013), peace and conflict studies (Crost, Felter, and Johnston 2014), and criminology (Chen and Shapiro 2007; Vollaard 2009).

But despite RD’s simple and intuitive nature, applications often face subtle problems that can pose serious threats to inference. The concern that tends to receive the most attention is that certain units close to the cut-point may be able to manipulate their scores and sort into the treatment or control groups (Imbens and Lemieux 2008; Caughey and Sekhon 2011; Eggers et al. 2014). However, even when units cannot sort, other complications can still occur that are just as problematic. These issues usually arise either from the complex rules of the scoring system or because researchers want to make certain comparisons that are not necessarily guaranteed to be valid by the RD.

While applied researchers have made important progress in addressing these problems, there are cases where they have confused the issues, sometimes simply because they wanted to avoid undesired complexity in their articles. No doubt, students hoping to learn
from past examples of RD could easily be misguided by the choices made in some of the most important applications. Unfortunately, there have only been a few attempts to clarify and resolve the complications that can occur in real world situations (Keele and Titiunik 2014), and some of the most widespread issues have not been explored in scholarly writing.

In this essay, I investigate three of the most important problems that arise in applied RD, all of which have been largely unaddressed by methodologists. The first occurs when units are organized into strata that vary in terms of their competitiveness. For instance, the units could be candidates running in different districts, and the districts could vary in terms of number of candidates with a realistic chance of winning. In these cases, researchers must be careful which units they include in their sample, or else they risk inducing bias. The second problem occurs when units appear in the sample multiple times, such as when students retake a test or candidates rerun for office. This scenario can cause some units to have multiple scores and possibly sort from one side of the cut-point to the other. The third arises when the independent variable of interest is not actually assigned by the RD. For instance, researchers might want to estimate how the likelihood of international conflict changes when a woman barely defeats a man in an election for head-of-state, even though gender is a basic human characteristic that was never itself randomized to politicians. This issue can make it difficult to interpret the treatment effect.

To help clarify these problems, I treat regression discontinuity as a natural experiment where the treatment is assigned randomly to units within a certain distance of the cut-point (Cattaneo, Frandsen, and Titiunik 2014). This approach differs from the more common practice of treating units’ scores as fixed and estimating the difference at the cut-point with two regression lines. However, in the next section I show that although these two approaches are conceptually very different, they are extremely similar mathematically. Their only major difference is that the continuity approach controls for the score. By doing so, the continuity approach has an important advantage. However, it can also make it harder to identify fundamental design problems that are easier to see when using the natural experiment approach. A major advantage of the natural experiment setup is that it prompts researchers to consider whether all units have roughly the same probability of treatment assignment close to the cut-point.

SECTION 1: THE BASICS OF REGRESSION DISCONTINUITY

Regression discontinuity is a quasi-experimental research design that estimates causal effects by exploiting scoring systems with important thresholds. The classic example is an exam where every student who scores above a cut-point receives an award and every students who scores below does not (Thistlethwaite and Campbell 1960). The idea is that receiving the award should be close to random for students who barely made or barely
missed the cutoff. Thus, the cut-point creates a source of exogeneity that gives researchers a natural experiment.

In many cases, all of the units that score above the threshold will receive the treatment, while all of the units that score below do not. These situations are referred to as sharp RDs. However, this clean set-up is not always necessary. We can still estimate the causal effects if some of the units that surpassed the cut-point did not receive the treatment or some units that fell short did. The key is that scoring above the cut-point cannot lower any unit’s probability of receiving the treatment, and it must increase the probability that at least some units got the treatment. In these situations, surpassing the cut-point can be thought of as receiving an instrument that increases the probability of treatment assignment for at least some units. These cases are called fuzzy RDs, and they can be analyzed using a combination of regression discontinuity and instrumental variable techniques.

To make this paper more accessible to readers, I write out the notation in terms of the conventional (sharp) RD design where all units to the right of the cut-point get the treatment and all units to the left do not. However, the main points of this paper carry over to fuzzy RDs with the intention-to-treat estimator, since all fuzzy RDs are sharp RDs if we take the instrument to be our treatment of interest.

There are two ways that researchers analyze regression discontinuities. I describe each of them below and then discuss their similarities and differences.

**The Natural Experiment Set-up.** This approach treats regression discontinuity as a natural experiment where the treatment is assigned randomly to units within a certain distance of the cut-point. Under this set-up, analysis is very straightforward. First, researchers must choose a size for their regression discontinuity window. Ideally, they make this decision prior to looking at the outcomes, and they are normally expected to report the results for other reasonable window choices as a robustness check. Next, researchers focus their attention on the units inside the RD window, excluding the others so that they do not affect the results. This move is valid even if it was somewhat random which units ended up in the RD window. It is similar to a situation where units were selected to be in an experiment in a semi-random manner. Which units ended up in the experiment could be taken as fixed, and then internally valid inferences could be drawn about that sample.

Next, the units inside the RD window are treated as though they were part of a real experiment. The standard notation is as follows. Denote the number of units as $n$. Unit $i$'s treatment status is $T_i \in \{0, 1\}$ and its score as $Z_i$. Under the assumption of non-interference between units, each unit has two potential outcomes, one under treatment ($Y_{it}$) and the other under control ($Y_{ic}$). The parameter that we are interested in is the Local Average Treatment Effect (LATE), which is written as

$$\bar{\tau}_{RD} = \frac{1}{n} \sum_i (Y_{it} - Y_{ic})$$

This parameter is just the average treatment effect for units inside the RD window. If $m$ is the number of treated units, then the estimator is
which is simply the difference in means between the treated and control units inside the RD window. As in a normal experiment, the standard error of this estimator is typically approximated by bootstrapping or by using the formula

$$SE = \sqrt{\frac{s^2_t}{m} + \frac{s^2_c}{n-m}}$$

where $s^2_t$ and $s^2_c$ are the estimated variances of the treated and control units. These are each calculated using the standard sample-to-population variance formula

$$s^2 = \frac{1}{N-1} \sum (x_i - \bar{x})^2$$

The p-value can be approximated using the estimated standard error, or it can be computed exactly (for the sharp null hypothesis of no treatment effect) by using permutation inference.

As with real experiments, it is usually possible to decrease the standard errors and reduce bias from baseline differences between the two groups by controlling for covariates that are predictive of the outcome. For instance, we could run a regression on the sample or do difference-in-differences estimation, as is commonly done with experimental data. However, these adjustment methods are usually considered robustness checks rather than valid procedures to get the main results. The reason is that allowing researchers to control for whatever covariates they want makes it possible for them to manipulate their results so that they can get lower p-values (Masicampo and Lalande 2012).

However, when analyzing a regression discontinuity, there is one additional reason to consider performing covariate adjustment that does apply to real experiments. In experiments, the treatment is randomized, making the unadjusted difference in means estimator is unbiased. On the other hand, regression discontinuity gives us two groups that were not actually randomized. For instance, if we are comparing leaders who won or lost their elections by less than 100 votes, we should expect the bare winners to differ from the bare losers in small but systematic ways. We might expect the bare winners to be slightly higher quality, wealthier, or smarter than the bare losers, and if we had a large enough sample size we would actually be able to pin-point these differences using difference in means tests. The most direct way to account for these systematic differences is to control for the score, $Z$, using some linear or non-parametric model. This step will eliminate small baseline imbalances that result from differences in the score for units on either side of the cut-point, provided that the relationship between the score and outcome is modeled correctly. I will discuss some possible models in the next section, as well as draw connections between these models and the second regression discontinuity approach.

**The Continuity Set-up.** Rather than treating regression discontinuity as a localized natural experiment, most researchers include all of the data and use two regression lines to estimate the difference between the average outcomes of treated and control units at the
cut-point. By using regression lines, they are treating the scores as fixed rather than random. The randomness is now in the error terms of the y’s, and the regression lines are the conditional mean functions that estimate the expected value of y across different values of the score. I call this approach the continuity set-up. When researchers use it, the major question that they face is how to construct the regression lines. Specifically, they have two key decisions to make.

One of these decisions is what bandwidth to use. The bandwidth, denoted as $h$, can be thought of as analogous to the size of the RD window. It specifies how far to the right or left of each point $z$ to look to compute the value of the regression line at $z$. For instance, if the bandwidth is set at $h = 2$, then the value of the regression function at $z$ would be estimated by using observations that lie between $(z-2, z+2)$. The bandwidth would normally be chosen using an optimal bandwidth selection algorithm like the one provided by Imbens and Kalyanaraman (2011) or Calonico, Cattaneo, and Titiunik (2014).

The other decision that researchers need to make is how to estimate the regression function using the points within the bandwidth. They could simply use the mean of these points (nearest-neighbor smoother) or a weighted mean that discounts observations the farther they are from $z$ (kernel smoother). They could also run a (new) regression line through these points and take the $\hat{y}$ value at $z$ (they would then repeat this process for every $z$ to construct their two regression lines). This method is called local regression. For instance, they could run weighted or unweighted regression through the points around in $(z-h, z+h)$ for every $z$ (local linear regression), or they could use polynomial regression (local polynomial regression).

But regardless of how they decide to construct their two regression lines, the regression discontinuity estimator is

$$\hat{t}_{RD} = E[Y|Z = c^+] - E[Y|Z = c^-]$$

which is just the difference between the two regression lines at the cut-point. Thus, when calculating their estimate, they are only focusing on the value of the regression lines at the cut-point, where $Z=c$. By doing so, they are ignoring all units that are farther than $h$ away from the cut-point. This is why the bandwidth is analogous to the size of the RD window. It specifies how far to the left and right of the cut-point researchers will look to find points to estimate the values of the two regression lines at the cut-point.

**Comparing the Two RD Approaches.** On a conceptual level, the two regression discontinuity approaches are very different. The natural experiment approach estimates the average treatment effect for the units within the RD window. This is a well defined causal effect provided that the treatment was as-if random for those units. On the other hand, the continuity approach estimates the average treatment effect at the cut-point, where no units exist. In this sense, there is no sample where this treatment effect is defined, making it radically different than a traditional causal parameter. Moreover, the continuity approach makes no as-if random assumption, except at the cut-point where there are no units.
Nonetheless, the two approaches are very similar mathematically. As mentioned before, the size of the RD window in the natural experiment approach is analogous to the bandwidth in the continuity approach. All points that fall outside the RD window (or bandwidth) have no impact on the results. In the same way, if we take the natural experiment approach and control for only the score, how we control for it is analogous to how we construct the regression lines in the continuity approach. In fact, as long as the bandwidth equals the size of the RD window and we use the same method to compute the standard errors, the natural experiment and continuity set-ups will have mathematically equivalent forms.

The parallels are very straightforward. If we take the natural experiment approach and conduct a weighted difference in means test that weights units by their scores, it is mathematically equivalent to using a kernel smoother in the continuity approach, provided we used the same kernel (or weighting rule) for both approaches. If we take the natural experiment approach and run a regression on the sample within the RD window that controls for \( Z \) and the interaction \( Z \times T \), that is equivalent to using local regression linear regression (unweighted) in the continuity approach. If we take the natural experiment approach and control for \( Z, Z^2, ... , Z^k \) and the interactions \( Z \times T, Z^2 \times T, ..., Z^k \times T \), that is equivalent to using a local polynomial regression of order \( k \) (unweighted) in the continuity approach.

Figure 1 summarizes these relationships graphically, using a study by Voeten (2013) that looks at how being elected to the UN Security Council influences countries’ willingness to participate in peace keeping missions. Voting for the UN Security Council is sometimes done through competitive elections, where countries secure a position if they receive at least 50% of the votes. I set the bandwidth (or size of the RD window) at \( h = 0.1 \), or 10%, which is close to the optimal bandwidth using the algorithm proposed by Calonico, Cattaneo, and Titiunik (2014). Thus, we are focusing on countries that received between 40% and 60% approval. Normally, we would also want to try smaller bandwidths where the as-if random assumption is more plausible, especially if we were not controlling for the score. However, since we are just using this study as an illustrative example, we will fix the bandwidth at 10%.

The figure shows that using the mathematical similarity between the natural experiment and continuity approaches. If we use the natural experiment approach and control for the systematic discrepancies between the treatment and control groups resulting from small differences in the score, we are really just doing a version of the continuity approach. Which version we are doing simply depends on how we control for the score.

A key point here is the importance of the as-if randomness for both approaches. All that the continuity set-up addresses is the bias that can result from units on one side of the cut-point being slightly different than units on the other side due to the small differences in their score. If certain units are more likely to be in bare winners than others for any reason other than the small differences in their scores, than the continuity approach is not longer unbiased. For instance, we can live with the fact that candidates that won by less than
Figure 1. Similarity of the RD Approaches

Natural Experiment Setup

Y = α + β₁T + β₂Z + β₃T * Z

\[ \tilde{\tau} = 0.4705 \quad p = 0.7 \]

Continuity Setup

Y = α + β₁T + β₂Z + β₃T * Z + β₄T * Z²

\[ \tilde{\tau} = -1.885 \quad p = 0.24 \]

Kernel Smoother

Local Linear Regression (Unweighted)

Local Polynomial Regression (2nd Order)

\[ \tilde{\tau} = -1.885 \quad p = 0.24 \]

Note: For the p-values to be equal, they must be bootstrapped, as there are small differences in some of the methods for computing p-values across the approaches.

100 votes might tend to be of slightly higher quality than candidates who lost by less than 100 votes. This bias can be eliminated by using the continuity approach, or by taking the natural experiment approach and controlling the score. However, we cannot live with the situation where certain units had other advantages that increased their chances of landing
in the bare winning group. Of course, if we control for the score, we must also assume that the model correctly captures the relationship between the score and the outcome.

Now while these two approaches are very similar mathematically, they each have their own advantages conceptually. In particular, the natural experiment approach is particularly useful for thinking about the design. Specifically, we can ask whether units that are very close to the cut-point all had roughly the same probability of being treated, or if some units had an advantage that cannot be accounted for by controlling for score. This question does not make sense in the continuity set-up, as each unit’s score is considered fixed. However, it is an essential question for any regression discontinuity design. It can help researchers identify major problems and their solutions, as I will illustrate in the next three sections.

**Section 2: Variation in Competitiveness Across Strata**

In the simplest RD set-up, units are all grouped together into the same pool, such as test-takers competing for a fixed number of scholarships. In these cases, there is only one strata, and all units belong to it. Analysis under this set-up is fairly straightforward. As long as the units around the cut-point could not manipulate their scores in a precise way, the treatment assignment around the cut-off should be “as-if” random. We should not expect imbalances on factors like gender, wealth, or ideology, since these should be balanced by the exogeneity of the treatment.

However, many regression discontinuities have more complicated formats where units are grouped into strata. For example, RDs that involve elections have units that are grouped into electoral strata (districts, states, countries, etc.) and the candidates in each strata compete against each other for office. This format has also appeared in other types of RD, including Van der Klaauw (2002), Niu, Xinchun, and Tienda (2013), and Bertoli (2015a).

The problem is that differences across strata could bias the results if they correlate with the probability of barely winning or barely losing. Before I lay out this problem formally, let me first illustrate it with an example. Consider an electoral system with two major parties. Now imagine that a third party decides to run in districts where it believes it has a good chance of winning, and that in many cases it scores close to the cut-point, along with the other two parties. For these districts where three candidates are in the RD window, the candidates would have roughly a 1/3 chance of barely winning and a 2/3 chance of barely losing. Thus, the candidates running in these districts would be more likely to be bare losers. If these candidates differed in systematic ways from the candidates running in districts with only two candidates in the RD window, the unequal probability of treatment assignment could bias the results.

This problem can arise anytime the strata vary in terms of their competitiveness. In these cases, more competitive strata will tend to have more bare losers than less competitive strata. In the election example, a district with four competitive parties may have three bare losers, whereas a district with two competitive parties can only have one bare loser. Thus,
when we combine a sample of bare winners and bare losers in close elections, we would
expect the following two observable implications:

1) The bare loser group should be larger than the bare winner group.

2) The bare loser group should have more units from competitive districts than the
bare-winner group.

We should also expect these problems to worsen as the number of districts with multiple
parties close to the cut-point grows.

This source of imbalance will be particularly dangerous if we are trying to estimate
something like the party incumbency advantage, which many recent studies have tried to
do in multi-party systems (Uppal 2009; Titiunik 2009). Bare losers will be more likely
to be running in districts where they are up against multiple rival parties that have a good
chance of winning. Thus, when we look at whether these parties won in their next election,
we should expect them to win at lower rates than if they were running against just one
competitive party. In short, they are more likely to lose this time and more likely to lose
next time. The result is that barely losing will appear to be more costly than it really is,
causing us to overestimate the incumbency advantage.

Researchers using the continuity approach might argue that their design avoids this
problem because it estimates the LATE exactly at the cut-point, and the probability of a
third place candidate existing at this point rapidly converges to 0 for a continuous run-
ning variable. Most importantly, it converges to 0 much faster than the probability for
second-place candidate being at the cut-point. In other words, while it is very unlikely that
one candidate will lose by exactly one vote, it is far less likely that two candidates from
the same district will both lose by exactly one vote. Thus, the influence of third-place
candidates should be negligible at the cut-point.

However, in RD applications, inferences about the cut-point are always drawn from the
data around the cut-point, and multiple bare losers from one strata can influence the results.
As I will show momentarily, whether controlling for $Z$ helps eliminate this bias depends on
whether the model correctly captures the way that the third-place units’ influence decreases
the closer they get to the cut-point. If it overestimates or underestimates the rate at which
their influence changes, then the regression function can be severely biased at the cut-point.

Before we get to that, let us first write this problem out formally using the simple natural
experiment set-up. The estimator we will use is $\hat{\tau}_{Full}$, which will be the difference in means
estimator for all units that fall in the RD window. We will assume for now that there can
only be one winner, as in electoral cases. Then our estimator is written as

$$\hat{\tau}_{Full} = \frac{1}{m} \sum_{i=1}^{n} Y_{it} T_{i} - \frac{1}{n-m} \sum_{i=1}^{n} Y_{ic} (1 - T_{i})$$
Let $C_i$ denote the number of units that are in the RD window in Unit $i$’s strata. So $I(C_i = k)$ is an indicator variable that equals 1 if there are $k$ units in the RD window in Unit $i$’s strata, and equals 0 otherwise. Moreover, let $q$ be the maximum number of units in the RD window in any strata. Then $\hat{\tau}_{Full}$ can be decomposed as follows

$$\hat{\tau}_{Full} = \frac{1}{m} \sum_{i=1}^{n} Y_{it} I(C_i = 2) - \frac{1}{n-m} \sum_{i=1}^{n} Y_{it} I(C_i = 2) +$$
$$\frac{1}{m} \sum_{i=1}^{n} Y_{it} I(C_i = 3) - \frac{1}{n-m} \sum_{i=1}^{n} Y_{it} I(C_i = 3) +$$
$$\cdot$$
$$\frac{1}{m} \sum_{i=1}^{n} Y_{it} I(C_i = q) - \frac{1}{n-m} \sum_{i=1}^{n} Y_{it} I(C_i = q)$$

Now note that for a strata with $k$ units in the RD window, the probability of barely winning is $P(T_i = 1) = \frac{1}{k}$ and the probability of barely losing is $P(T_i = 0) = \frac{k-1}{k}$. Using these expression, we can write out the expected value of our estimator as follows:

$$E[\hat{\tau}_{Full}] = \frac{1}{m} \cdot \frac{1}{2} \sum_{i=1}^{n} Y_{it} I(C_i = 2) - \frac{1}{n-m} \cdot \frac{1}{2} \sum_{i=1}^{n} Y_{it} I(C_i = 2) +$$
$$\frac{1}{m} \cdot \frac{1}{3} \sum_{i=1}^{n} Y_{it} I(C_i = 3) - \frac{1}{n-m} \cdot \frac{2}{3} \sum_{i=1}^{n} Y_{it} I(C_i = 3) +$$
$$\cdot$$
$$\frac{1}{m} \cdot \frac{1}{q} \sum_{i=1}^{n} Y_{it} I(C_i = q) - \frac{1}{n-m} \cdot \frac{q-1}{q} \sum_{i=1}^{n} Y_{it} I(C_i = q)$$

which reduces to

$$E[\hat{\tau}_{Full}] = \frac{1}{n} \sum_{i=1}^{n} [Y_{it} I(C_i = 2) + Y_{it} I(C_i = 3)/(3/2) + \ldots + Y_{it} I(C_i = k)/(k/2)] -$$
$$\frac{1}{n} \sum_{i=1}^{n} [Y_{it} I(C_i = 2) - (3/2) \cdot Y_{it} I(C_i = 3) - (k/2) \cdot Y_{it} I(C_i = k)]$$

Of the two terms on the right hand side, the top estimates the average outcome under treatment, $\frac{1}{n} \sum_{i=1}^{n} Y_{it}$, and the bottom estimates the average outcome under control, $\frac{1}{n} \sum_{i=1}^{n} Y_{ic}$. However, the estimate of the average outcome under treatment underweights units from strata with many units in the RD window, while the estimate for the average outcome under control overweights these units.

Now let us examine this problem by looking at legislative elections across a number countries. Eggers et al. (2014) compile a dataset that provides the vote shares of the first three parties in about 200,000 elections across ten countries, which include two-party and multi-party systems. In two-party systems, there are only two candidates close to the cut-point in any election, aside from the rare cases where an independent or minor third-party candidate performs very well. Thus, we should not expect variation in competitiveness across strata to be a major problem when dealing with two-party systems. Without this variation, there should not be many more bare losers than there are bare winners. On the hand, in countries where there are several competitive parties, we should expect third-party candidates to cause differences in competitiveness across strata, which will lead to a larger bare loser group.
Figure 2 shows whether there tended to be more bare losers in the countries using the McCrory sorting test. This test checks for whether there was an overall tendency for units to sort to one side of the cut-point or the other. Any such sorting would be a threat to inference, unless it was entirely independent of the potential outcomes. As this figure shows, countries that had more districts with multiple candidates near the cut-point tended to fail the McCrory sorting test. Specifically, they had more bare losers than we would expect if each candidate was about equally likely to be on one side of the cut-point or the other.

Figure 3 shows that countries that had more districts with multiple candidates near the cut-point also tended to have more imbalance on the competitiveness of district. There
is no imbalance for countries like Australia, France, and New Zealand, where there are only two competitive parties in close elections. However, the p-values are close to 0 for counties like India, Mexico, Brazil, and the United Kingdom (which has the largest number of close elections aside from the United States). For the entire sample, about 2.8% of the bare losing candidates came from competitive districts, compared to about 1.4% of the bare winning candidates. For a t-test in the RD window, the p-value is $p \approx 1.7 \cdot 10^{-11}$. 

Note: p-values values are computed using difference in means tests for the sample of candidates that won or lost by less than 2%.
In fact, it is fairly easy to see that this imbalance must exist. Each winner has exactly one corresponding runner-up. For instance, if the winning party won by 1%, then the runner-up must have lost by 1%. Moreover, both of these units are either from a district with only two parties close to the cut-point, or they are both from a more competitive district with three parties close to the cut-point. Thus, the winners and runner-ups by themselves will always be balanced on competitiveness of district, since they are a mirror image of each other. However, in cases where there is a third-place party in the RD window, that party adds an additional unit in that district. When this happens in many districts, these additional bare losers will create the imbalance that is evident in Figure 3.

Because the density of third-place candidates decreases the closer we get to the cut-point, our estimator might be unbiased if we control for $Z$ with a model that accurately captures the decreasing influence of third place candidates as they approach the cut-point. In the case of elections with three parties, the influence of third-place parties will actually decrease linearly close to the cut-point. For example, imagine that we set the size of the RD window at (-2%,+2%). The score of a third-place candidate in this RD window is (roughly) the minimum of two draws from the random variable Unif(-2%,0%), which has a linear density (in this case, $f(x) = -x/2$ for $x \in (-2,0)$)). Thus, local linear regression can substantially reduce bias in the three-party case. The kernel smoother, however, can lead to badly biased results, as it incorrectly models the relationship between $Z$ and competitiveness. The nearest-neighbor smoother can also be biased, since it is equivalent to a difference in means test.

While the local linear smoother can help us avoid bias when some strata have three candidates in the RD window, it cannot resolve cases where some strata have four or more candidates close to the cut-point. Put simply, these strata will induce a non-linear relationship between $Z$ and the influence of additional bare losers. The fourth place candidates will be distributed $f_1(x) = 3x^2/8$ and the third place candidates will be distributed for $f_2(x) = -3x^2/2 - 3x$ for $x \in (-2,0)$. These distributions are easily derived using order statistics formulas. Without the linearly decreasing influence of the additional bare losers, the local linear smoother can be badly biased. For instance, if you take Eggers et al. dataset of roughly 20,000 bare winners and bare losers and add just 91 fourth-place candidates, each of whom scores half-way between -2% and the third-place candidates, the local linear estimator returns significant imbalance on competitiveness of district. While high-order polynomial smoothers may be able to capture the non-linear influence of the additional bare losers to some extent, they will also be more sensitive to noise in the data and have a much higher variance (Gelman and Imbens 2014).

In short, when strata have varying numbers of units in the RD window, how we control for $Z$ matters. It is no longer a robustness check, but a key decision about the design that can greatly impact the results. In the three-candidate case, we can go from no bias with a local linear smoother to substantial bias with a kernel or nearest-neighbor smoother. If we add some fourth-place candidates into the dataset, local linear regression will also
become biased. Put simply, the additional bare losers make our results highly sensitive to our modeling assumptions.

Since one of the most appealing features of RD is that it tends to be much less sensitive to how we model the data, we should consider some ways of resolving this problem. The easiest solution is just to restrict the analysis to the two units in each strata that were closest to making and closest to missing the cut-point. While this approach involves dropping units, and therefore losing some information, it maintains balance over strata and time. In other words, the bare winners and bare losers in the election example will be perfectly balanced on district and year, since every bare winner has a corresponding bare loser. Thus, this procedure creates a blocking scheme that should increase balance in the sample. Researchers can also still adjust for $Z$ and present the normal continuity graphs with two regression lines that meet at the cut-point. It is just that all units that were not the closest to making or missing the cut-point would be left out of this graph.

The second solution is to reweight units based on their probability of treatment assignment given their strata. If Unit $i$ is from a strata with $k$ units in the RD window and $j$ bare winners, then we would give Unit $i$ weight $\frac{k}{n} \cdot \frac{k-j}{k}$ if it barely won and weight $\frac{k}{n} \cdot \frac{j}{k}$ if it barely lost. We could then use a difference in means test under the natural experiment set-up. In electoral settings, this estimator would be

$$\hat{t}_{ReweightedFull} = \frac{1}{n_2/2} \sum_{i=1}^{n_2} Y_{it} I(C_i = 2) - \frac{1}{n_2/2} \sum_{i=1}^{n_2} Y_{ic} (1 - T_i) I(C_i = 2) +$$

$$\frac{1}{n_3/3} \sum_{i=1}^{n_3} Y_{it} I(C_i = 3) - \frac{1}{2n_3/3} \sum_{i=1}^{n_3} Y_{ic} (1 - T_i) I(C_i = 3) +$$

$$\frac{1}{n_q/q} \sum_{i=1}^{n_q} Y_{it} I(C_i = q) - \frac{1}{(q-1)n_q/q} \sum_{i=1}^{n_q} Y_{ic} (1 - T_i) I(C_i = q)$$

where $n_k$ is the number of units from strata with $k$ outcomes close to the cut-point. This estimator is easy proven to be unbiased under the as-if random assumption by taking the expected value of $\hat{t}_{ReweightedFull}$. We could still control for the score by using weighted regression with this weighting scheme, thus getting the benefits of local linear or local polynomial regression. By reweighting, this procedure maintains balance across strata while also using all of the data in the RD window. The drawback is that researchers can no longer construct the continuity graph, since variation in the units’ weights cannot be easily represented graphically.

**SECTION 3: UNITS THAT REAPPEAR IN THE SAMPLE**

There are two versions of the multiple-score problem. The simple version occurs when units have outcomes for each scoring round, such that each score is associated with a single outcome. In these cases, researchers can simply treat every time that a unit receives a score as a separate observation. The more problematic version arises when each unit has only one outcome, meaning that multiple scores will now correspond to single outcome.
For instance, if we wanted to estimate how passing an exam influenced the chances that students went to college, we would have a problem if the students could take the exam multiple times. However, there is a fairly straightforward solution to this problem that I will discuss in the second half of this section.

**Unique Outcomes for Every Score.** This issue arises for many studies, including (Lee 2008), Hainmueller and Kern (2008), Lalive (2008), Lemieux and Milligan (2008), Broockman (2009), Titiunik (2009), Cellini, Ferreira, and Rothstein (2010), Hall (2014), Eggers et al. (2014), and Bertoli (2015a; 2015b). For example, consider the RD studies that look at the incumbency advantage. Their goal is to estimate how winning an election at time $t$ affects a party’s vote share and winning probability at time $t + 1$. To increase the sample size, these studies look at long historical periods rather than a single election year, treating each election as a separate observation. The units are parties within each district. These parties’ treatment assignment is whether they won, their score is how close they were to winning, and their outcome is how they did in the next election. One entry in the dataset would be Party $k$ in District $i$ with the treatment assigned at time $t$ and the outcome assigned at time $t + 1$, another would be the same party and district with the treatment assigned at $t + 1$ and the outcome assigned at $t + 2$, and so on. Thus, the problem of having multiple scores for single outcomes does not arise, but there is still the issue of rerunning units.

In these cases, the key question is whether the treatment effect in each round influences the LATE in future rounds. If there is good reason to believe that it does not, then researchers can consider each case where a unit received a score can as a single observation. However, if the treatment assignment in one round affects either who is in the RD window later or the size of future individual treatment effects, then researchers need to be more careful about how they proceed.

Since this issue is probably clearest in the party incumbency advantage literature, I will stick with that example. First, let us assume that treatment assignment is as good as random within a small window around the cut-point. Now imagine that winning an election at time $t$ increases vote share at $t + 1$ by 3 points. This implies that treatment assignment in round $t$ can affect which units are very close to the cut-point at time $t + 1$, so treatment in early rounds affects who is influencing the LATE in later rounds. Furthermore, imagine that some units also receive a 5-point bump for winning if they won in the previous round, but they only experience a 3-point bump if they lost in the previous round. If some of these units are in the RD window back-to-back years, then treatment assignment affects both who influences the LATE and the size of treatment effects for some units after the first round.

We can still test the sharp null hypothesis that incumbency has no impact of vote share in future rounds, since this hypothesis guarantees independence between rounds. However, deriving point estimates and confidence intervals for the LATE is no longer a simple matter. On one hand, there is no longer a single LATE, since both who is close to the
cut-point and how large some of the individual treatment effects are in later rounds depend on treatment assignment in earlier rounds. Thus, the LATE is a random variable. The parameter that we might be interested in estimating is the expected value of the LATE over all treatment assignments:

$$\text{E} \left[ \text{LATE} \right] = \text{E} \left[ \frac{n_1}{n} \text{LATE}_1 \right] + \text{E} \left[ \frac{n_2}{n} \text{LATE}_2 \right] + \ldots + \text{E} \left[ \frac{n_k}{n} \text{LATE}_k \right]$$

Surprisingly, the normal RD procedures give us an unbiased estimate of this value, assuming that the treatment is random within the RD window and we restrict our attention to the pair of candidates closest to the cut-point. Let $T$ be the treatment assignment vector, which is composed of the treatment assignment vectors in each of the $k$ rounds.

$$T = T_1 \cup T_2 \cup \ldots \cup T_k$$

Let the number of units per round be denoted by $n_1, n_2, \ldots, n_k$ units, which are also random variables that depend on the treatment assignments in previous rounds, with the exception of $n_1$. Furthermore, the LATE is a random variable that is the sum of the LATE’s from each round, $\text{LATE}_1, \text{LATE}_2, \ldots, \text{LATE}_k$, where each of these terms is multiplied by the percentage of units that are coming from that round:

$$\text{LATE} = \frac{n_1}{n} \text{LATE}_1 + \frac{n_2}{n} \text{LATE}_2 + \ldots + \frac{n_k}{n} \text{LATE}_k$$

So the parameter we are interested in is

$$E[\text{LATE}] = E \left[ \frac{n_1}{n} \text{LATE}_1 \right] + E \left[ \frac{n_2}{n} \text{LATE}_2 \right] + \ldots + E \left[ \frac{n_k}{n} \text{LATE}_k \right]$$

The normal RD estimator can also be broken down into the estimates for each round. The estimate for the first round is clearly unbiased under the normal as-if randomness assumption, since there is no rerunning problem in this round. However, $T_1$ is not only one of $2^{n_1}$ possible treatment assignments for Round 1, but also one of $2^{n_2}$ possible paths to Round 2, which is selected at random. Moreover, the RD estimate for Round 2 is an unbiased estimator for all possible treatment assignments in Round 2 on that path. So we randomly selected our path to Round 2, and then got an unbiased estimate for Round 2 conditional on that path. Similarly, $T_2$ is a randomly selected path to Round 3, and we again get an unbiased estimate of the LATE in that round conditional on that path. This pattern continues until we reach the last round.

The result is that the full path from Round 1 to Round $k$ is randomly selected from all possible paths, and in each round along the way we have an unbiased estimator for the LATE in that round conditional on being on that path. So we have

$$E[\tau_{\text{RD}}] = E \left[ \frac{n_1}{n} \text{LATE}_1 \right] + E \left[ \frac{n_2}{n} \text{LATE}_2 | T_1 \right] + \ldots + E \left[ \frac{n_k}{n} \text{LATE}_k | T_{k-1}, \ldots, T_1 \right]$$

The tower property gives us

$$E[\tau_{\text{RD}}] = E \left[ \frac{n_1}{n} \text{LATE}_1 \right] + E \left[ \frac{n_2}{n} \text{LATE}_2 \right] + \ldots + E \left[ \frac{n_k}{n} \text{LATE}_k \right]$$

$$E[\tau_{\text{RD}}] = E[\text{LATE}]$$
So even when the size of the effects and observations in the RD window depend on the treatment, we can still estimate an important parameter of interest. It should be noted, however, that we do not have an unbiased estimate of the realized LATE that occurred on the observed path unconditional on the path, unless that LATE equals $E[LATE]$. It will also not necessarily be an unbiased estimate of the realized LATE conditional on the path. For instance, some LATE’s may only come about if the bare winners and losers in the first round are very different on baseline covariates, and in these cases $\tau_{RD}$ will be biased from the chance imbalance early on.

While we can obtain an unbiased point estimate of $E[LATE]$, confidence intervals are more complicated? Unfortunately, we have no way of determining what the distribution of the LATE looks like. We only have unbiased estimates for each round on a randomly determined path, but with possible correlation between treatment assignment and the LATE in future rounds. Simulations suggest that the normal RD confidence intervals can be misleading if the LATE in future rounds is highly dependent on treatment assignment in earlier rounds. Of course, this problem does not apply if we are simply testing the sharp null hypothesis.

This issue is a major limitation if we care about quantifying the uncertainty of our point estimate, which is certainly true in the party incumbency literature. In cases like this, where we both care about the confidence interval and believe that treatment assignment influences the LATE in later rounds, it would probably make more sense to take a different approach. One option would be to drop every district after it appears in the RD window once, which would limit the sample size and change the parameter to a different LATE, but would provide us with a valid confidence interval under the normal RD assumptions.

---

**Only One Outcome for Each Unit.** A more complicated scenario arises when units have multiple scores but only one outcome. In these cases, there are usually units that both make and fall short of the cut-point at different times. The problem is that it is not entirely clear whether they belong in the treatment or control group, or what values of their scores should be used.

The most famous application that deals with this problem Eggers and Hainmueller (2009) APSR article “MP’s for Sale: Returns to Office in Postwar British Politics,” which tests how winning a seat in British Parliament influences wealth at death. To get around the rerunning issue, they count all candidates who won at least once as winners and use their scores from their first winning races, while counting all candidates who never won as losers and using their scores from their best losing races. For instance, a candidate who ran twice and lost with scores $\{-30\%, -2\%\}$ would be given the score $-2\%$, and a candidate who ran four times with scores $\{-5\%, 14\%, -1\%, -4\%\}$ would be given the first winning score of 14%. Eggers and Hainmueller then use the usual RD procedures on this new sample, where each candidate has only one score and outcome. I call the estimator that is obtained by these procedures the First-Winning-Best-Losing (FWBL) estimator.
While the FWBL estimator ensures that no unit will have multiple scores, it creates several other problems. The first is that the expected size of the treatment group will now be much larger than the expected size of the control group. This problem occurs because bare losers can switch to winners by rerunning, but bare winners are immediately fixed as bare winners provided that they did not already win in a previous round.

Thus, there are four types of candidates: (1) candidates who lose on their first attempts and fall outside the RD window, and who can rerun in later rounds and move right; (2) candidates who lose on their first attempt and fall inside the RD window, and who can also rerun in later rounds and move right; (3) candidates who win on their first attempts and fall inside the RD window, and who now have fixed scores in the treatment group; and (4) candidates who win on their first attempts and fall outside the RD window, whose scores are now also fixed at their first winning score. Although the sizes of the bare winner and bare loser groups should be about equal after every candidate’s first run, candidates in the second group (bare losers) can rerun and switch to the bare winner group or to the group of winners who scored outside the RD window. The result is that the bare loser group gets depleted, with some of their members switching to the bare winning group. While initial losers in the first group can rerun and replace them, these units are just as likely to be bare winners as bare losers in any future round. Moreover, even if they move up to become bare losers, they can still sort out of this group by winning an election in a future round.

In short, the more opportunities that candidates have to win, the more likely they are to be counted as a bare winners and the less likely they are to be counted as bare losers. Thus, within the RD window, treatment should be correlated with the likelihood of rerunning. More specifically, it should be correlated with their probabilities of rerunning after losing, since they become fixed in the sample after their first win. Since a candidate’s probability of rerunning after losing is likely to be related to individual resources and determination, there is reason to suspect that the FWBL estimator will be biased in many cases.

There is also another potential source of bias that would arise even if all units were guaranteed to run the same number of times. Imagine that there are two types of candidates, those who respond to barely losing by working hard and succeeding later, and those who get frustrated and put little effort into future campaigns. The more resilient candidates would tend to sort out of the bare loser category in later rounds, whereas their more easily discouraged counterparts would lose badly and be counted as bare losers based on their performance in the first round. Thus, even if there was a predetermined rule that required each candidate to run exactly two times, the FWBL could be confounded by this additional source of bias. Specifically, we would expect bare losers to be less resilient than bare winners.

The nature of this bias is fundamentally more problematic than the imbalance on competitiveness across strata in Section 2 because a well-chosen smoother cannot eliminate this type of bias. In the prior example, the influence of the additional bare losers converged to 0 at the cut-point, since the probability of multiple candidates from one strata...
barely losing rapidly converged to 0 when $Z$ is continuous. Thus, we could eliminate bias if we chose a smoother that accurately captured the relationship between $Z$ and the decreasing influence of the additional bare losers. However, in this case the influence of candidates rerunning does not converge to 0 the closer we get to the cut-point. Candidates that lose by one vote would probably be very likely to rerun, and would bias the FWBL estimator if they won in a later election.

One potential solution to this problem is to use a different estimator that is proposed by Querubin and Snyder (2011) in a similar study. Their study looks at how winning a seat in US Congress affected individual wealth during the mid-nineteenth century. Rather than looking at each candidates first win or best loss, they look at the first time each candidate ran. Thus, they count any candidate who lost in his first round as a loser, even if he won and held office in a later round. Their justification for using this estimator is that only about 9% of candidates who lost on their first try attempted to rerun, so very few of their losers ended up holding office later.

This estimator can be thought of as an intention-to-treat (ITT) estimator. The idea here is that a candidate’s outcome on his first try is an instrument for whether he ever held office. Winning on the first attempt guarantees that the candidate would hold office, whereas losing increases the probability that the candidate would never hold office. Thus, we can think of the candidates who lost on their first attempt and never held office as compliers, and the candidates who lost on their first attempt but won later as non-compliers. Of course, all units who won on their first attempt are compliers, since they were guaranteed to hold office. Thus, we have an instrument (winning or losing on the first attempt) and one-way non-compliance, making the estimator that Querubin and Snyder use an ITT estimator.

A third approach is just to acknowledge that the outcome in the first round is an instrument and proceed with the fuzzy regression discontinuity design. Fuzzy RDs are situations where the cut-point determines who receives an instrument instead of who receives the treatment. The Fuzzy RD estimator is the ITT estimator divided by the estimated local compliance rate, which is defined as the percentage of first-time losing candidates in the RD window who never held office. I denote this parameter as $\alpha_{LATE}$. So the estimator is

$$LATE_{FRD} = \frac{LATE_{ITT}}{\alpha_{LATE}}$$
Since $\alpha_{LATE} < 1$, this value will have a larger absolute value than the ITT estimator. As with any fuzzy RD, the p-value for this estimator is the same as the ITT estimator p-value, which tests whether the instrument has an affect on the outcome.

The difference between the ITT and Fuzzy RD estimators is that the ITT estimator captures the impact of the instrument (succeeding the first time), whereas the Fuzzy RD estimator captures the effect of the treatment (succeeding at some point). In most applications, the Fuzzy RD estimator will probably be preferable, even when the compliance rate is high. Since the p-values from the two approaches are the same, the decision to use one estimator over the other matters for the size of the effect but not the statistical significance of the results.

There are two drawbacks of the Fuzzy RD estimator. First, it requires researchers to assume that the exclusion restriction holds. The instrument cannot affect the outcome through in any way aside from affecting the likelihood of the treatment. This assumption may be invalid in many cases. In the economic returns to office example, we would have to assume that the candidates’ outcome in their first races do not affect their future wealth except in how they change the probability of holding office at some point. This assumption would be violated if there were candidates who would win a future election if they barely fail the first time, and whether they succeeded the first time affects their future wealth. In this scenario, their performance on the first run does not affect their probability of holding office, which is 1 regardless of whether they win or lose. However, it does influence their outcomes, thus violating the exclusion restriction. It is therefore very important that researchers consider whether the exclusion restriction holds before using the Fuzzy RD estimator.

Second, the Fuzzy RD estimator will usually be biased in finite samples, as is normally the case with instrumental variable estimators. The reason is that we can estimate both $LATE_{ITT}$ and $\alpha_{LATE}$ without bias, but we cannot convert these values into an unbiased estimate of $\frac{LATE_{ITT}}{\alpha_{LATE}}$. This problem results from the rule in probability that $E[A] E[B] \neq E[AB]$ in general. However, if the sample size were to go to infinity, the variances of the numerator and denominator would go to 0, giving us an unbiased estimate of the $\frac{LATE_{ITT}}{\alpha_{LATE}}$. Therefore, the Fuzzy RD estimator is consistent, albeit not unbiased.

It is also possible to increase the power of both of the ITT and Fuzzy RD estimators by disregarding any attempt where a candidate scored below the RD window. In other words, only count an attempt as an attempt if the unit won or scored within the RD window. Thus, any early attempt where a candidate scored below the RD window would not disqualify that unit from the analysis. This rule will increase the sample size by retaining any candidate who had a bad early loss but produced a more competitive score later. I call these estimators the Refined ITT and Refined Fuzzy RD estimators.
In Figure 1, I compare the five estimators using a simple simulation. The simulation procedures are as follows, although the patterns observed here do not depend on the specific details, but will arise anytime the probability of rerunning is correlated with baseline determinants of the outcome. There are four election years, each with 10,000 candidates who run against unnamed opponents. Candidates who score above 50% win their elections, and candidates that score below 50% lose. Each candidate has a quality level $Q_i \sim \text{Unif}(0.2, 0.8)$. The vote share for a candidate in any election is $Z_i = Q_i + \varepsilon$, where $\varepsilon \sim \text{Unif}(-0.2, 0.2)$. Every candidate reruns with probability $P_i \sim \text{Unif}(0, 1)$, and candidates that drop out are replaced by new candidates. $T_i$ is the treatment indicator which equals 1 if the candidate ever held office and 0 if not. A candidate’s baseline wealth is $B_i = 100,000 \cdot P_i + \varepsilon^*$, where $\varepsilon^*$ is another random error term. Finally, wealth at death is $W_i = B_i + 10,000 \cdot T_i$. Thus, the treatment effect of holding office is a constant of $10,000$, and wealthier candidates are slightly more likely to rerun than poorer candidates.

The left-hand column shows the density of the running variable for each estimator. All densities are smooth around the cut-point except for the FWBL estimator, which we expected to be discontinuous since bare losers can sort right. The middle column shows the bias for each estimator as a function of the correlation between the baseline outcome and the probability of rerunning. As expected, the FWBL estimator is substantially biased, since wealthier candidates are more likely to rerun and become bare winners if we use their first-winning or best-losing vote share. The Fuzzy RD and Refined Fuzzy RD estimators are also biased, even though the exclusion restriction holds in this example. The other two estimators perform very well, although they focus on the intention to treat effect (or the effect of winning the first time). The third column presents the standard errors for each estimator. The standard errors for the Refined ITT and Fuzzy RD and estimators are slightly lower than the standard errors for the unrefined estimators, since these estimators increase the sample size.

It should be noted, however, that this simulation may not capture what is happening in Hainmueller’s and Eggers’s study. It depends on the extent to which the probability of rerunning correlates with wealth, and they only have a measure of wealth at the time of each candidate’s death. Querubin and Snyder do have an estimate of baseline wealth. However, very few of their candidates ever won after losing the first-time, so the compliance rate is too large to distinguish between these four types of estimators using their data. Thus, it is difficult to illustrate this problem with a real-world example. However, there are clear theoretical reasons to believe that the FWBL estimator will provide misleading results in cases where the compliance rate is lower (many units win after their first loss) and their is a strong correlation between the outcome and the probability of rerunning, and the simulation illustrates the problem very clearly.
Figure 4: Comparing Estimators with a Simulation

Note: The dotted horizontal line in the middle graph shows the intention-to-treat effect. Thus, the Refined Intention-to-Treat Estimator provides the best estimate of the intention-to-treat effect, and the Refined Fuzzy RD estimator provides the best estimate of the treatment effect.
SECTION 4: INDEPENDENT VARIABLES NOT ASSIGNED BY THE RD

In the previous examples, scoring above or below the cut-point determined who received some treatment, like holding office or earning a scholarship. The main question of interest was how that treatment affected units, which could be estimated if the treatment was as-if random around the cut-point. However, there are other cases where the main variable that researchers care about is not actually a treatment, but an attribute. For instance, Meyersson (2014) compares Turkish municipalities where an Islamic party barely won and barely lost elections to test how Islamic rule affects women’s rights. Other scholars have focused on cases where Democrats or Republicans narrowly won or lost gubernatorial races, looking at outcomes like taxation (Fredriksson, Wang, and Warren 2013), unemployment (Leigh 2007), and racial inequality (Beland 2014). Another example is Clots-Figuerasa’s (2012) use of close races where women barely won or lost their elections to test whether female politicians increase the likelihood that children will receive a primary education.

This type of regression discontinuity can be very informative, but it is first important to recognize what we cannot learn from it using the normal RD assumptions alone. That is, it does not allow us to estimate the impact of the attribute of interest on the outcome. For instance, Meyersson finds that the municipalities where an Islamist party barely won had better women’s rights records afterward, but this finding tells us little about the effect of political Islam on women’s rights. After all, Islamist parties differ from the non-Islamist parties in many ways besides religion and ideology, and the observed treatment effect could be due to some of these other differences. Put another way, his study tells us very little about how convincing a secular politician to join an Islamist party would affect women’s rights. Similarly, Leigh finds that electing Democratic governors tends to result in less unemployment, but this does not mean that if we convinced a Republican governor to become a Democrat we should expect a similar effect. After all, Republican and Democratic governors differ in many ways besides their ideologies.

To understand what this type of regression discontinuity tells us under the normal assumptions, we must think of the RD set-up in a different way than before. In the more traditional RD cases, like when test-takers try to earn a scholarship, the units are the individuals and the treatment is getting the scholarship. However, in cases where the independent variable of interest is an attribute, the unit is no longer the individual. Instead, the units are the open “spots” that individuals or parties can fill, such as seats in Congress. The treatment is whether that open spot received an individual with a particular characteristic, although that characteristic was not randomly assigned across individuals or parties and is likely confounded with other important factors.

In this context, the regression discontinuity gives us causal leverage by creating exchangeability across the open spots, rather than across the units that fill them. For instance, municipalities where Islamist parties barely defeated secular parties should be similar to municipalities where they barely lost to secular parties. This exchangeability allows us
to be confident that there are not systematic differences between districts controlled by Islamist and secular parties, but we still must be cautious that differences between the parties besides political Islam may be driving the results.

Thus, this type of regression discontinuity tells us how the outcome will tend to change if the seat to be occupied by someone of Type A or Type B, given that the race is tight. For example, if we were voters in Turkey who cared about women’s rights, we should consider voting for the Islamist party if the election was a toss-up. However, whether the causal mechanism has anything to do with the party being Islamist is an open question. Similarly, Clots-Figuera’s study tells us whether voting for a female politician in a close election is good for educational levels. However, it tells us very little about the effect of gender on education, since the female politicians in her study are not being compared to a similar group of counterfactual male candidates. We simply do not know what that comparison might reveal. Therefore, this type of regression discontinuity is more useful for predicting the impact of a certain decision, in these cases who to elect in tight races, than it is for understanding the effect of characteristics that were never assigned by the RD.

However, the RD can make it easier to estimated the effect of the characteristic when combined with observational methods. What the RD does is guarantee that the environments where units of Type A win are comparable to the environments where units of Type B win. If we can control for some of the confounding factors that make candidates of Type A different than candidates of Type B, then we may have a better case for arguing differences in outcomes are explained by the characteristics.

Consider the best case scenario where we randomly assigned certain politicians to be Republicans or Democrats prior to the election. If we observed that economic outcomes were better when the Republicans won office, we could not be sure that result was explained by our randomized treatment. After all, Republicans might be elected more often when the economy is already doing well, meaning that systematic differences in the external environment are confounding our analysis. However, if we randomized both party identification and who won the elections, we would have an unbiased estimate of the impact of our treatment on the economic outcomes.

**CONCLUSION**

This paper has covered several key points, which can be summarized as follows:

(1) The natural experiment approach and continuity approach are mathematically similar despite their conceptual differences. The question is not about which of these approaches researchers use, but how they control for the score. A major advantage of the natural experiment approach is that it can helps researchers evaluate their research designs.
(2) In cases where units are grouped together into strata that vary in terms of their competitiveness, researchers should restrict their focus to the pair of units that are closest to the cut-point—the highest scoring loser and the lowest scoring winner. Failing to do so can result in bias if the smoother does not accurately capture the relationship between $Z$ and the changing influence of additional bare winners and losers on the estimates.

(3) Units that are in the sample multiple times do not pose a problem if they have one outcome for every score, although they can pose a threat to quantifying the uncertainty of the estimates. This problem can be managed by dropping certain units from the sample.

(4) When units with multiple scores have only one outcome, researchers should use the Refined Fuzzy RD design, where the instrument is whether the unit won or lost in the first case where it scored in the RD window or above it. It is important to consider whether the exclusion restriction holds in these cases.

(5) RDs cannot help researchers estimate the impact of a characteristic that was not assigned by the RD. They can only help researchers determine what will happen if an individual with that characteristic barely wins or barely loses.

Keeping these points in mind can help researchers avoid bias in applications of RD.

The subtle problems explored in this paper also underscore how important it is for researchers to offer a clear theoretical explanation for why the treatment should be as-if random around the cut-point. Because of the emphasis on balance tests, not much attention is payed to whether the design makes sense in theory. The key is to show that all units close to the cut-point should have roughly the same probability of treatment assignment. The only possible systematic difference would result from the small imbalance in the score, and that could be accounted for by controlling for it.

If researchers take advantage of the natural experiment approach to identify potential problems with their designs, then RD can be a very powerful tool for inference. Scoring system are a central feature of modern society, used in areas like education, sports, public health, criminal justice, domestic politics, and international relations. Although the scoring system will often be more complicated than the classic example of high-school students taking a test, researchers who are careful about their designs can use these cases to make new and fascinating discoveries.
REFERENCES

Bernardi, Fabrizio and Emmanuel Skoufias. 2014. “Compensatory Advantage as a Mechanism of Educational Inequality A Regression Discontinuity Based on Month of Birth.” Sociology of Education 87(2) 74-88.


