# Effect Heterogeneity and Causal Attribution in Regression Discontinuty Designs: Introducing the Moderation-in-Discontinuities Framework *

Kirk Bansak[†] Tobias Nowacki[‡]

January 2022

Abstract

Research investigating subgroup differences in treatment effects, recovered using regression discontinuity (RD) designs has become increasingly popular. For instance, 'difference-in-discontinuity' designs assess whether incumbency effects on candidate persistence or winning again vary by candidate characteristics (e.g., gender) or local context. Under what conditions can we interpret this subgroup difference in treatment effects as a causal effect of the moderating characteristic? In this paper, we explicitly explore the difference between RD effect conditionality that is simply associated with versus causally driven by other variables. We introduce a method for the estimation of causal moderation effects in RD settings and formally define an alternative 'moderation-in-discontinuity' estimand with a causal interpretation. Next, we introduce estimators under selection-on-observable-type identification assumptions. We apply our framework to three recent studies that investigate differences in conditional RD treatment effects and offer 'best practice' advice for applied researchers, highlighting the importance of carefully interpreting the research design's quantity of interest. Our paper contributes to the literature on conditional RD treatment effects and offers a new estimation strategy suited towards answering important questions about causal moderation effects in electoral settings and beyond.

Keywords: regression discontinuity, conditional treatment effect, moderation effect, political methodology

*Acknowledgments: We thank Dan Thompson and participants at the 2021 Stanford Causal Science Center Conference for helpful feedback.

[†]Assistant Professor, Department of Political Science, University of California San Diego, 9500 Gilman Drive, La Jolla, CA 92093, United States. E-mail: kbansak@ucsd.edu

[‡]PhD Candidate, Department of Political Science, Stanford University, 616 Jane Stanford Way, Stanford, CA 94305, United States. E-mail: tnowacki@stanford.edu

# 1. INTRODUCTION

Originally proposed by Thistlethwaite and Campbell (1960), the regression discontinuity (RD) design has in recent years become one of the most popular, established methods used by social scientists for investigating causal effects. Often described as a form of "natural experiment" (Dunning, 2012; Titiunik, 2021), the RD design takes advantage of situations in which the delivery/receipt of an intervention or treatment depends upon whether a unit is above or below a predetermined (and often arbitrary) threshold on an observed covariate. Such situations exist across a wide variety of domains in human society, where decisions must often be made on the basis of administrative thresholds with respect to continuous variables, such as age, vote percentages, group or population sizes, etc.). Under certain conditions, the causal effect of the intervention or treatment can then be reliably estimated by comparing the trends of units close to and on either side of the threshold. In addition to the prevalence of situations that fit into the RD framework, the RD design's popularity has also been aided by evidence of its high internal validity, recovering estimates close to benchmarks from randomized experiments (e.g. Chaplin et al., 2018; Cook and Wong, 2008).

In political science, the RD design has in particular been used in electoral scholarship, as winning an election depends on whether or not a candidate's or party's vote share exceeds a predetermined threshold. This has allowed for political scientists to estimate the causal effects of candidate or party incumbency on reelection and a range of other downstream outcomes (Eggers et al., 2015). The RD design has also been used by political scientists to study various other phenomena as well (Abou-Chadi and Krause, 2020; Anagol and Fujiwara, 2016; Dunaiski, 2021; Eggers, 2015; Fujiwara, 2011; Gerber et al., 2011; Kaplan et al., 2019; Pettersson-Lidbom, 2008; Yoder et al., 2021).

As usage of the RD design has become more pervasive across the social sciences, with researchers often iterating in greater nuance and detail on topics investigated by prior scholarship, the analysis of heterogeneity in RD-based causal effects has also become increasingly popular. (Anderson, 2014; Barrow et al., 2020; Card and Giuliano, 2016; Hansen, 2015; Jenkins et al., 2016) Such investigations, often presented under the label "difference-in-discontinuities",[1] consider whether RD-based causal effects vary as a function of other characteristics or across different subsets of their populations of interest. This strand of research mirrors a similar focus that has expanded in experimental research on conditional average treatment effects and treatment effect heterogeneity (see e.g. Gerber and Green, 2012; Imai

[1]Though note that the term "difference-in-discontinuities" has been used by different scholars to describe several slightly different designs, a point that will be revisited later.

and Ratkovic, 2013; Ratkovic, 2021; Ratkovic and Tingley, 2017; Wager and Athey, 2018).

Along these lines in political science, a number of electoral RD studies have investigated how the effect of incumbency varies conditional upon the characteristics of the candidate, party, or local context (Bernhard and de Benedictis-Kessner, 2021; de Benedictis-Kessner, 2018; Eggers and Spirling, 2017; Lopes da Fonseca, 2017; Núñez, 2018; Olson, 2020; Wasserman, 2018, 2021). For instance, Wasserman (2018) finds a stark gender gap in candidates' persistence after losing an election at the local level in California: losing an election decreases the chance of running again by 50% for women than it does for men, although the gender gap may attenuate in more senior electoral settings (Bernhard and de Benedictis-Kessner, 2021; Wasserman, 2021). Similarly, Brown et al. (2020) offer evidence that women are less likely to rise through the ranks of the office pipeline: the effect of being elected to the state legislature on their probability of ever winning a seat in Congress is far smaller than the equivalent for men.

However, once there is evidence of conditionality of the RD effect, interpreting the nature of that conditionality is a separate matter. Specifically, the question is whether one can attribute any RD effect differences across subgroups to the causal influence of the conditioning variables themselves. For instance, in the case of different incumbency effects across gender, what is the actual *causal* role of gender? Is the difference in the RD effect simply associated with gender, or it is causally due to gender (or voters' perceptions of gender)? This is a distinction that is important from both policy and theoretical perspectives. For instance, the causal source of the incumbency effect gender gap (and whether or not it is due to gender itself) has implications for how widely it would manifest across different contexts. Indeed, while Wasserman (2018) finds a gender gap in candidates' post-loss persistence at the local level, both Bernhard and de Benedictis-Kessner (2021) and Wasserman (2021) find gender parity in the effect of losing at the level of statewide offices. These divergent results provide suggestive evidence that any difference in conditional effects across gender may be due not to gender alone, but also the election context and correlated candidate attributes (e.g., experience); Wasserman (2021) suggests that her statewide and local results are different because candidates at the state level have more experience. However, rigorous assessment of this possibility would require specialized analysis.

Unfortunately, studies that investigate patterns of conditionality in RD effects are often unclear on (a) whether or not their underlying claim is that the conditionality is actually causally driven by the conditioning variable, and (b) whether or not the research design and estimation procedures employed can actually provide reliable evidence to that effect.

It appears that a key barrier in this regard is not simply the conceptual complexity of distinguishing between the different types of conditionality, but also the challenges of knowing how to formalize and estimate each type and under what conditions that can (or cannot) be done. As a result, the nature of the conditionality of RD effects is often discussed and analyzed by applied scholars in ways that are confusing and/or not sufficiently explicit, making it difficult to know what final conclusions can be drawn. In some studies, the theoretical role of the conditioning variable may not be explicitly discussed at all, leaving the reader to wonder what the precise quantity or phenomenon of interest is. In other studies, the theoretical motivation for investigating a particular moderator variable involves a specific mechanism or more general process that implies a causal influence of that moderator variable, but the empirical analyses are undertaken in such a way that suggests otherwise, leading to a confusing mismatch between theory and empirics. In yet other studies, the authors may be clear about their quantity of interest, and they may theoretically motivate their empirical strategy, but insufficient formalization is provided to convincingly establish that the research design and estimation procedures do indeed map onto their quantity of interest. In sum, the norms and practices currently employed when researchers discuss and investigate RD effect conditionality are prone to confusion.

This paper seeks to help address these issues, allowing applied scholars investigating conditional RD effects to be more explicit in their quantities of interest, more intentional in their empirical analyses, and more precise in the interpretation of their results and the resulting theoretical and/or policy implications. We explicitly explore the difference between RD effect conditionality that is simply associated with vs. causally driven by other variables. To make this distinction explicit and formal, we define two alternative estimands that capture this difference within the potential outcomes framework. The first, which we term the *Heterogeneity-in-Discontinuities*, refers to the difference in RD effects across observed levels of some other variable. The second, which we term *Moderation-in-Discontinuities*, refers to the difference in RD effects that is actually caused by (in a counterfactual sense) the the other variable. We lay out different sets of identification assumptions that are required for each estimand, along with corresponding estimation procedures. In doing so, we emphasize the additional challenges involved in both identification and estimation of the Moderation-in-Discontinuities estimand, highlighting the carefulness and caution that applied researchers must take in interpreting evidence of RD effect conditionality. As part of this, we discuss key considerations and present diagnostics for considering the plausibility of the identification assumptions behind the Moderation-in-Discontinuities estimand. To illustrate the methods

and further highlight the challenges involved, we present X applications with varying degrees of plausibility that Moderation-in-Discontinuities is identified. We conclude that it is possible to generate reasonable evidence of Moderation-in-Discontinuities in some contexts, but that in other contexts there is little to no plausibility of doing so, and applied researchers need to be more careful in their analysis and interpretation of RD effect conditionality.

## 2. MOTIVATION AND RELATED WORK

### 2.1. Motivating Examples

Both the Heterogeneity-in-Discontinuities and the Moderation-in-Discontinuities estimand can be important and policy-relevant quantities – depending on the research question and theorized mechanism at hand. Yet, researchers are, unfortunately, often imprecise in distinguishing between the two estimands in their interpretation of results. In other cases, even where practitioners are more careful, we might care about a more robust causal interpretation of the conditioning variable, as it pertains to the policy-relevant question. In this section, we provide (two) motivating examples that highlight why we consider this distinction critical.

First, consider the role of gender in conditioning the effect of losing on candidates' decision to run again (Cipullo, 2021; Wasserman, 2018). We want to understand whether women are differentially less likely to run again if they lose an election – any such gender difference might contribute to the persistent underrepresentation of women in politics. Multiple mechanisms can account for a differentially greater attrition among women – not all of which are causally derived from gender. Women might be less likely to run again due to voter bias, or more negative experiences during their first campaign. In these cases, politicians' gender is the causal moderator of the difference in persistence. Equally plausibly, however, women might be less likely to run again if, on average, they are older, or if they have more precarious careers (from which they can afford less time off). In such cases, the causal mechanism originates from sources correlated with, but distinct from gender, even though women running again less frequently is an observable consequence.

Understanding the root cause of the gender differential is crucial for policy implications: if being a women has a true, *causal* effect on candidate persistence, policymakers may want to address this issue by explicitly promoting more women as candidates (or maybe this would imply that gender quotas are more effective). If, on the other hand, the true conditioning effect results from a variable that is simply correlated with gender – for example, age or prior experience – policymakers may want to consider alternative ways of addressing the problem. That said, the descriptive difference in RD effects may still be interesting and important

(e.g., to evaluate downstream consequences of diminished female representation). In sum, we may have reason to care about both the descriptive heterogeneity in effects as well as the causally interpretable moderation in effects. Crucially, the two quantities are important for different reasons, and so we need to be careful in distinguishing them.

Another application of Heterogeneity-in-Discontinuities studies how the timing and nature of elections affect the magnitude of incumbency advantages: we want to understand whether particular electoral environments can diminish electoral accountability.We might infer from the difference in incumbency advantages across parties (Eggers, 2015), election cycles (de Benedictis-Kessner, 2018) or primary regimes (Olson, 2020) that some elected officeholders enjoy greater electoral safety once elected (implying potentially diminished accountability, or allowing legislators to extend their time horizon). Here, too, the distinction between the two quantities of interest is critical for our understanding of why this difference persists, and informs possible implications for policymakers: If reformers wished to address this discrepancy, we need to know whether the difference in effects is *caused* by the highlighted characteristic, or merely a correlational byproduct of some other mechanism – for example, certain parties or off-cycle elections attracting lower-quality candidates.

In most empirical work to date, this important distinction between the two quantities of interest is not as clearly made as it can (and should) be. Unsurprisingly, there is wide heterogeneity in how carefully researchers distinguish between the descriptive difference in effects between subgroups, and a causal interpretation of said difference. Often, while authors refrain from direct causal claims, the explication is ambiguous or omits the distinction. Consider, for example Bohlken (2018):

> Table 3 examines whether the occurrence of co-partisan elite targeting after a state election depends on whether the MP belongs to the state ruling party (i.e., the party of her state's chief minister).

Here, the interpretation of results doest not explicitly attribute a causal role to the moderator (MP belongs to the state ruling party), but the use of 'depends' also implies a role for the moderator beyond heterogeneous treatment effects. Such ambiguity is, unfortunately, common, and aggravated by the lack of a unified formal framework to discuss RD effect conditionality.

Even where scholars diligently stress that the conditioning variable of interest may be correlated with other factors, they do not offer a formalization for the identification assumptions necessary to recover a causal interpretation. Consider, for example, the following two quotes from de Benedictis-Kessner (2018) and Wasserman (2018), respectively:

However, these results are only a descriptive difference between those elections that are on-cycle and those that are off-cycle, and subject to some degree of selection bias. For instance, cities that have on-cycle elections might also just elect worse-quality incumbents than cities with off-cycle elections. These incumbents might then perform worse in their subsequent elections. Additionally, cities that have on-cycle elections might have different features than those cities with off-cycle elections. Some cities even determine their own election timing, and the factors influencing this decision might be the cause of these differences in the incumbency advantage. Identifying the true effect of £election timing on the incumbency advantage requires circumventing this selection problem.

Although RD relies on the continuity of candidate covariates throughout the threshold for winning, the identifying assumption does not ensure covariate continuity across subgroups. This means that among close losing candidates, male and female candidates may have different characteristics that can potentially contribute to their divergent responses.

These examples illustrate the importance of carefully distinguishing between Heterogeneity-in-Discontinuities and Moderation-in-Discontinuities estimands. Heterogeneity-in-discontinuiies designs are widely used and easily specified. In the remainder of the paper, we offer a new framework for identifying and estimating Moderation-in-Discontinuities designs.

### 2.2. "Difference-in-Discontinuities" and its Variants

The term "difference-in-discontinuities" has often been employed by researchers investigating RD effect conditionality. However, usage of this term has been highly variable—an example of the inconsistent norms and practices in this area. The term has alternatively been used to describe either an estimand or an estimation strategy/approach, with studies often not being explicit or clear about this distinction. Even more confusingly, the term has been used to refer to entirely different underlying estimands of interest. The following provides a lengthy discussion of the term difference-in-discontinuities, characterizes (and distinguishes between) the different ways in which the term has been used, and makes explicit how the variants of the difference-in-discontinuities relate to the quantities of interest in this study.

The first way in which the term difference-in-discontinuities has been used deals with the investigation of the difference in RD effects across two (or more) subsets of the data or population of interest. This relates directly to the investigation of treatment effect heterogeneity or conditional average treatment effects in regression discontinuity designs (e.g.

Becker et al., 2013; Hsu and Shen, 2019); it is essentially comparing multiple, separate conditional RD effects. Most frequently, this looks at the difference in RD effects across subsets of the population as defined by particular background characteristics (e.g. Card and Giuliano, 2016; Desai and Frey, 2021; Lalive, 2007; **?**). Additionally, this framework has also been used to investigate the difference in RD effects across different contexts (e.g. to help unpack compound treatments) (Card and Giuliano, 2016). This usage of the term corresponds to the Heterogeneity-in-Discontinuities estimand presented and formalized in the present study.

A second way in which the term difference-in-discontinuities has been used builds on the first, but uses time as the conditioning variable. That is, some studies have estimated a change in an RD effect of interest across two time periods (Chicoine, 2017; De Benedetto and De Paola, 2019; Grembi et al., 2016; Kantorowicz and Köppl–Turyna, 2019; Köppl-Turyna and Kantorowicz, 2020), where those time periods are associated with some other change/policy intervention. (These studies rely on estimating the effect at the discontinuity in two different time periods, because in $t = 0$, observations on either side of the threshold may still have a baseline difference in $Y_0$ if things other than the intended treatment change at the discontinuity. ) At first appearances, this is in principle the same as in the first case, as the only difference is the specific variable that is being used as the conditioner. However, it is worthwhile to distinguish this from the first because, when doing this, such studies have different motivations than simply seeking heterogeneity of the RD effect across time. Instead, some studies applying this approach have sought to imply that that change/policy actually has a causal influence on the RD effect (cites). This, of course, would only be the case if the only thing that changed over the two time periods was the policy, so such claims are not necessarily immediately plausible. In other cases, however, while the estimator may be this variant of a "difference-in-discontinuities," the estimand may be something entirely different. For instance, this approach has been paired with additional assumptions such that the estimator instead identifies a causal effect of the policy variable itself that is changing over the two periods (Grembi et al., 2016).

Finally, a third way in which the term difference-in-discontinuities has been used refers to attempts to incorporate a more extensive panel data structure, along the lines of generalized difference-in-differences methods, into the regression discontinuity design. The motivation behind leveraging time more richly in this manner is precisely to attempt to estimate a causal effect of some other variable on the RD effect. However, in practice, the ways in which this has been done has lacked standardization and formal clarity, with the difference-in-differences elements layered into the RD specification being motivated more by intuition

and heuristic rather than a solidly rigorous formalization. This is not entirely surprising. For one, as will be shown below, there are important subtleties in formalizing the causal effect of some variable on an RD effect even in the absence of a panel structure—and in the absence of needing to make parallel trends of other difference-in-differences-type assumptions. Second, generalized difference-in-differences and panel event modeling is an area of research that is currently active, fast-changing, and (as recent papers have shown) sufficiently complicated on its own without being further integrated with another identification strategy like RD.

This paper will distinguish between two fundamentally different types of RD effect conditionality, focusing its attention specifically on quantities of interest that do not build around or rely on a panel data structure. In doing so, we will be attentive to rigorously defining these different estimands and distinguishing those estimands from estimation approaches. First, the term Heterogeneity-in-Discontinuities will refer to an estimand that captures a difference in RD effects that manifests across values of a conditioning variable, but where there is no presumed causal influence of the conditioning variable on the RD effect. (This is along the lines of the first way described above in which the term difference-in-discontinuities has been used.) Second, this paper will present an alternative estimand termed Moderation-in-Discontinuities that specifically focuses on the causal influence of the conditioning variable on the RD effect—i.e. allowing for heterogeneity in an RD effect across a conditioning variable to be causally attributed to that conditioning variable. As mentioned above, both estimands will be defined explicitly outside of the panel context, though note that this does not completely preclude the use of time as a variable in the identification strategies and estimation procedures. Specifically, time may be considered as either the primary conditioning variable or as part of a covariate adjustment strategy. However, this paper does not focus on using time to take advantage of panel-data features and assumptions, such as parallel trends, as in the third variant of "difference-in-discontinuities" described in the previous paragraph.

### 2.3. Other Related Work

The present study also relates to other methodological and applied research using the RD design. As has already been alluded to, and as will be formalized later, the quantities of interest in this study entail the use of covariates, in addition to the forcing variable. Hence, there are links to previous methodological research on the inclusion of covariates in RD designs (Calonico et al., 2019; Frölich and Huber, 2019). However, these studies focus on the inclusion of covariates for the purposes of increasing the statistical precision of the the RD effect estimator, as well as to aid in achieving identification in the RD design (e.g. to control

for discontinuities in the potential outcomes or the covariate distribution at the cutoff). The use of covariates proposed and formalized in these studies contrasts with the present purposes. Here, it is assumed that conditions are met such that the aggregate RD effect of interest itself is identified (and estimable) without covariate adjustment, and the inclusion of covariates is not motivated by considerations of statistical precision. Instead, covariates are employed to investigate, identify, and estimate different forms of RD effect conditionality, as will be explained in greater formality in the next section.

Another related line of research has investigated the use of the RD design to specifically identify and estimate the effects of other variables (i.e. variables other than the treatment defined by the running variable). This has been particularly popular in election RD studies, with the strategy of leveraging close elections to estimate at the cutpoint the causal effects of race, partisan affiliation, and other characteristics of winning candidates on various downstream outcomes (for an overview, see Marshall, 2019). However, the validity of this strategy has been called into question by Marshall (2019), who highlights the inferential and estimation challenges of isolating the causal effect of a variable at the cutpoint, especially when that variable can affect the running variable. While the underlying quantity of interest in that line of research is different from the focus here, the Moderation-in-Discontinuities estimand that will be presented is defined in such a way that implicitly interfaces with and takes into account the issues articulated by Marshall (2019). Specifically, the estimand will formally allow for the conditioning variable of interest to have an effect on the running variable.

In another relevant piece of research, Feigenbaum et al. (2017) propose a clever approach using a multidimensional regression discontinuity design with an eye toward identification and estimation of the causal effect of a conditioning variable (majority-party status) on an RD effect (incumbency advantage). Hence, their theoretical quantity of interest relates closely in spirit with the Moderation-in-Discontinuities estimand that will be defined here. However, the links are much less close from the perspective of formalization, and the approach they propose appears to be highly specific to the context of studying majority-party control, rather than serving as a more general framework.

In another relevant piece of research, Jenkins et al. (2016) combine propensity scores with a design focused on RD effect heterogeneity. As they explain, the "PS weights induce comparability between Head Start and OK pre-K children [(i.e. units across each level of the conditioning covariate)], allowing us to make a statistical comparison of the two treatment effects in the same RD model." On the one hand, this strategy is strongly suggestive of an interest in the causal influence of the conditioning variable on the RD effect. On the other

hand, however, while the econometric specification is explicated, the precise estimand of interest is less clear and never formalized. Nonetheless, the estimation strategy appears to very much be in the spirit of what we call Moderation-in-Discontinuities below, as will be developed and further explained in detail later.

## 3. NOTATION AND SETUP

As in previous formalizations of the regression discontinuity design (e.g. Imbens and Lemieux, 2008), this study employs the potential outcomes framework of Neyman (1923) and Rubin (1974) to define causal effects. The previous scholarship has focused on potential outcomes defined with respect to a treatment, specifically $Y_i(t)$, with $t \in \{0, 1\}$ representing possible values of the treatment. In contrast, here we posit for each unit $i$ the existence of $Y_i(t, s) \in \mathbb{R}$ for $t \in \{0, 1\}$ and $s \in \mathcal{S}$, where $s$ represents the possible values of a third (pre-treatment) variable, which is the primary conditioning variable of interest and which we will call the moderator. $\mathcal{S}$ denotes its support. The potential outcome $Y_i(t, s)$ denotes the outcome unit $i$ would experience if that unit had treatment status $t$ and moderator value $s$. For simplicity, let $\mathcal{S} \equiv \{0, 1\}$, yielding potential outcomes $Y_i(0, 0), Y_i(1, 0), Y_i(0, 1), Y_i(1, 1)$.

Further, for each unit let $Y_i \in \mathbb{R}$ denote the observed outcome, where the relationship between the observed outcome and potential outcomes is governed by $Y_i = Y_i(T_i, S_i)$. As in the standard sharp regression discontinuity context, the observed treatment $T_i \in \{0, 1\}$ is assigned on the basis of a particular pre-treatment covariate (i.e. the running or forcing variable) denoted here by $X_i \in \mathbb{R}$, such that $T_i = \mathbf{1}(\{X_i > c\})$ for some cutoff $c \in \mathbb{R}$. In addition, $\boldsymbol{W}_i \in \boldsymbol{\mathcal{W}}$ denotes a vector of other observed pre-treatment covariates, which we will refer to as the "control set." Further, $S_i \in \{0, 1\}$ denotes the observed value of another pre-treatment covariate of focal interest, termed the moderator as described already above, which is believed to have some relationship with the causal effect of the treatment on the outcome.

Finally, precision in this formalization also requires considering the relationship between $S$ and $X$. Investigations of RD effect conditionality typically proceed upon the (often implicit) premise that the conditioning variable is prior to or not otherwise affected by the running variable—such analyses would otherwise imply conditioning on a post-treatment variable. We will also follow this practice and rule out the possibility of a causal effect of $X$ on $S$. However, applied research is often ambiguous about whether or not $S$ affects $X$, which can lead to confusion and difficulty in interpreting results (for a similar argument, see Marshall, 2019). Here, we will explicitly allow for the possibility that the moderator $S$ has a causal

effect on the running variable $X$. Formally, this implies the existence of $X_i(s) \in \mathbb{R}$ for $s \in \{0, 1\}$. Similar to the case of the potential outcomes $Y_i(t, s)$, the relationship between the observed $X_i$ and the counterfactual $X_i(s)$ is governed by $X_i = X_i(S_i)$. Note that this setup also nests the special case where there is no causal relationship between $S$ and $X$, in which case $X_i(0) = X_i(1) = X_i \ \forall \ i$.

With these definitions in place, we posit a data-generating distribution on the tuples $(Y_i(0,0), Y_i(1,0), Y_i(0,1), Y_i(1,1), X_i(0), X_i(1), S_i, \boldsymbol{W}_i)$. Now, suppose we observe $i = 1, ..., N$ independent and identically distributed samples of the form $(Y_i, X_i, S_i, \boldsymbol{W}_i)$. For each unit $i$, a tuple is drawn from the aforementioned distribution. As described above, for any unit $i$, $T_i = \mathbf{1}(\{X_i > c\})$, and the observed $Y_i$ and $X_i$ are determined by $Y_i(T_i, S_i)$ and $X_i(S_i)$, respectively.

## 4. HETEROGENEITY-IN-DISCONTINUITIES

### 4.1. Estimand

Using the notation above, the first estimand that we define is what we call the Heterogeneity-in-Discontinuities:

Definition 1 (Heterogeneity-in-Discontinuities)

$$E[Y_i(1, S_i) - Y_i(0, S_i)|X(S_i) = c, S_i = 1] - E[Y_i(1, S_i) - Y_i(0, S_i)|X(S_i) = c, S_i = 0]$$

$$= E[Y_i(1, 1) - Y_i(0, 1)|X = c, S_i = 1] - E[Y_i(1, 0) - Y_i(0, 0)|X = c, S_i = 0]$$

Since the potential outcomes in this estimand employ only the observed values $S_i$ for $s$, this estimand could equivalently be accommodated by the traditional RD design notation in which potential outcomes are defined only with respect to the treatment as $Y_i(t)$. In that case, this "reduced form" version of the estimand would be $E[Y_i(1) - Y_i(0)|X = c, S_i = 1] - E[Y_i(1) - Y_i(0)|X = c, S_i = 0]$. (The reason the definition is not presented exclusively in this reduced form, however, is to explicitly highlight the role that $S$ does (and does not play) and to draw a more direct comparison with the Moderation-in-Discontinuities estimand that will be presented later.)

The Heterogeneity-in-Discontinuities estimand falls directly under the umbrella of what has been discussed as treatment effect heterogeneity or conditional average treatment effects in regression discontinuity designs (Hsu and Shen, 2019). For applied researchers conducting analysis of conditional RD effects by estimating separate RD effects across different subsets, the most common route for analyzing RD effect conditionality in applied work, this analysis

maps onto the Heterogeneity-in-Discontinuities—whether the researcher knows/intends this or not.

### 4.2. Identification and Estimation

Identification of the Heterogeneity-in-Discontinuities can proceed from the standard RD formulation attributed to Hahn et al. (2001) (see also Imbens and Lemieux, 2008), which rests on the assumption of continuous conditional expectation functions through the cutpoint. Adapted to the present context, the necessary assumption is as follows:

Assumption 1 (Subset Continuity of Conditional Expectation Functions)

$$E[Y_i(t, S_i)|X_i = x, S_i = s]$$

*are continuous in* $x$ *for all* $t \in \{0, 1\}$ *and* $s \in \{0, 1\}$.

As can be seen, this continuity assumption proceeds by conditioning on $S$ and is limited to continuity in the expectation of potential outcomes only for the observed values of $S$ upon which the conditioning takes place. As such, it is analogous to the continuity assumption applied in the standard RD design, but simply applying that assumption to two separate subsets (defined by observable values of $S$).

Under this assumption, for any $s \in \{0, 1\}$:

$$E[Y_i(1, S_i)|X_i = c, S_i = s] = \lim_{x \downarrow c} E[Y_i(1, s)|X_i = x, S_i = s] = \lim_{x \downarrow c} E[Y_i|X_i = x, S_i = s]$$

And similarly, again for any $s \in \{0, 1\}$:

$$E[Y_i(0, S_i)|X_i = c, S_i = s] = \lim_{x \uparrow c} E[Y_i|X_i = x, S_i = s]$$

Hence, the Heterogeneity-in-Discontinuities estimand is identified as such:

$$\begin{aligned} & E[Y_i(1, S_i) - Y_i(0, S_i)|X = c, S_i = 1] - E[Y_i(1, S_i) - Y_i(0, S_i)|X = c, S_i = 0] \\ = & \left( \lim_{x \downarrow c} E[Y_i|X_i = x, S_i = 1] - \lim_{x \uparrow c} E[Y_i|X_i = x, S_i = 1] \right) \\ - & \left( \lim_{x \downarrow c} E[Y_i|X_i = x, S_i = 0] - \lim_{x \uparrow c} E[Y_i|X_i = x, S_i = 0] \right) \end{aligned} \tag{1}$$

This is tantamount to identifying two separate RDD-treatment effects—one for units with $S_i = 1$ and one for units with $S_i = 0$—and then taking their difference.

Similarly, estimating the Heterogeneity-in-Discontinuities estimand can proceed following normal RD estimation strategies, but applied separately for each subset $S_i = 0$ and $S_i = 1$.

For instance, a regression model can be employed, and the estimand can in fact be recovered by a single regression model for $E[Y_i|S_i, X_i]$. Working within the predominant RD estimation framework employed by applied researchers, adapting standard approaches to specifically recover the Heterogeneity-in-Discontinuities quantity leads to the following model, where the expected relationship between the outcome and running variable is allowed to vary on each side of the cutoff according to best practices in RD modeling more generally:

$$\begin{aligned} E[Y_i|S_i, X_i] &= \alpha_0 + \alpha_1\tilde{X}_i + \alpha_2 S_i + \alpha_3\tilde{X}_i \cdot S_i + \\ &\quad \beta_0 T_i + \beta_1 T_i \cdot \tilde{X}_i + \beta_2 T_i \cdot S_i + \beta_3 T_i \cdot \tilde{X}_i \cdot S_i \end{aligned}$$

where $\tilde{X}_i = X_i - c$. This model can then be estimated using a local linear regression (e.g. fitting this regression for data within a certain bandwidth of $X$, if using a local linear regression with a rectangular kernel). Under this model, $\beta_2$ identifies the Heterogeneity-in-Discontinuities estimand of interest. Note that this is mathematically equivalent to specifying two analogous standard RD regression models separately for each subset $S_i = 0$ and $S_i = 1$, and then taking the difference between their RD effects.

## 5. MODERATION-IN-DISCONTINUITIES

In contrast to the Heterogeneity-in-Discontinuities estimand's focus on differences in the RD effect (of the treatment on the outcome) across observed values of the moderator, one might instead want to know the difference in RD effects that might result by counterfactual intervention upon $S_i$. In other words, rather than knowing whether a difference in RD effects simply manifests observationally across values of $S_i$, the question is whether $S_i$ itself causes a difference in RD effects. And whereas there is a reduced form version of the Heterogeneity-in-Discontinuities estimand that makes use of the simple potential outcomes defined only with respect to $t$, the expanded potential outcomes defined with respect to both $t$ and $s$ are critical for understanding and formalizing this alternative phenomenon, which we term the Moderation-in-Discontinuities. As will be seen throughout this section, there are substantial additional complications involved in both identifying and estimating the Moderation-in-Discontinuities, relative to the Heterogeneity-in-Discontinuities.

### 5.1. Estimand

The Moderation-in-Discontinuities is defined as follows:

Definition 2 (Moderation-in-Discontinuities)

$$E[Y_i(1,1) - Y_i(0,1)|X(1) = c] - E[Y_i(1,0) - Y_i(0,0)|X(0) = c]$$

For the Moderation-in-Discontinuities to be a relevant and interesting estimand, it must be that the moderator of interest is itself mutable at the level of individual observations. This is in contrast to the Heterogeneity-in-Discontinuities estimand, and it should be one of the first considerations for applied researchers when deciding on their target estimand. It is also important to consider that the mutability of a particular variable or characteristic also depends upon how an observation is defined. For example, it may be argued that racial identity is a characteristic that is not sufficiently open to mutability, thereby hindering analysis of the causal dynamics of race from a counterfactual perspective. This may be the case if each unit of observation is a particular individual person. However, racial identity is clearly mutable for other units of observation—such as individual résumés, as in audit experiments, or individual electoral contests, where it is no strain on one imagination to consider a counterfactual election with a candidate that was similar in all ways except of a different race.

As already noted above, this setup allows for the possibility that $S$ affects $X$, which is why the expressions condition on $X(1)$ and $X(0)$. This estimand captures how the effect of the treatment on the outcome *at the cutpoint* $c$ would change if all units took moderator value $S = 1$ vs. $S = 0$. Of course, in the case where $S$ affects $X$, because different counterfactual values of the moderator also imply different values of $X$, the units that have a value of $X$ within any specified distance from $c$ will not necessarily be the same under different counterfactual values of $S$. Hence, it important to interpret the Moderation-in-Discontinuities as a moderation to the effect at the cutpoint (i.e. the discontinuous jump in the CEFs), not the moderation of an effect for any individual particular observation.[2] While this may at first impression seem odd, note that in the case of the more intuitive Heterogeneity-in-Discontinuities estimand, the effect of the treatment on the outcome is being compared for entirely different sets of units altogether. Further, this is a natural extension of the localness of a normal RD effect—that is, it captures the effect of the treatment on the outcome *at the cutpoint* $c$, which is not the effect for any particular observation and would not necessarily be the same under some other value of the cutpoint.[3]

In addition, note that the definition also covers the special case where there is no causal re-

[2]This also relates to the observations made by Marshall (2019), who highlights the challenges of identifying the causal effect of $S$ itself on the outcome at the cutpoint if $S$ also has a causal effect on the running variable. The formalization of the Moderation-in-Discontinuities estimand presented here hence implicitly echoes Marshall's concerns.

[3]To further consider the implications as in a normal RDD, and whether or one can extrapolate any findings beyond the cutpoint $c$, one must consider how different are the slopes of the potential outcomes in expectation on the left and right for each moderator value.

lationship between $S$ and $X$. In that case, the estimand simplifies to $E[Y_i(1,1) - Y_i(0,1)|X = c] - E[Y_i(1,0) - Y_i(0,0)|X = c]$.

### 5.2. Identification and Estimation

Identification of the Moderation-in-Discontinuities estimand requires additional and more stringent assumptions than the Heterogeneity-in-Discontinuities estimand. In addition, there are a number of different identification strategies with alternate sets of assumptions that could be applied, each proceeding from different statistical design/model setups and implying different estimation procedures. The following section will begin by describing a strategy to identify and estimate the Moderation-in-Discontinuities when the moderator has been randomized (or can otherwise be assumed to be unconditionally exogenous). Note that in most applications, it may be unlikely that one's moderator of interest is random. Nonetheless, this setting is an important starting point for understanding the differences (and connections) between Moderation-in-Discontinuities and Heterogeneity-in-Discontinuities. It will also serve as an illuminating reference point for the Moderation-in-Discontinuities strategy presented in the section immediately following, which does not rely on a random moderator.

#### 5.2.1. Strategy with Exogenous Moderator

Similar to the identification of the Heterogeneity-in-Discontinuities presented above, identification of the Moderation-in-Discontinuities also follows the standard RD formulation based upon continuous conditional expectation functions. However, a slightly different version of continuity is required here, represented in the following assumption:

Assumption 2 (Full Continuity of Conditional Expectation Functions)

$$E[Y_i(t,s)|X_i(s) = x]$$

*are continuous in* $x$ *for all* $t \in \{0,1\}$ *and* $s \in \{0,1\}$.

Note that this continuity assumption is technically more stringent than the version of continuity required for identification of Heterogeneity-in-Discontinuities represented in Assumption 1. As described earlier, the Heterogeneity-in-Discontinuities' version of continuity is analogous to the continuity assumption applied in the standard RD design, but simply applying that assumption different subsets defined by observable values of $S$. In contrast, the continuity assumption required for the Moderation-in-Discontinuities is with respect to the expectation of all potential outcomes $Y(t,s)$, unconditional upon observable values of $S$. At

the same time, however, it is unclear that the version of continuity required for Moderation-in-Discontinuities is actually more demanding as a practical matter, as discussed in more detail later in Section 6.3.

In addition, this strategy proceeds by positing the exogeneity of the moderator (most plausible when the moderator has been explicitly randomized). Specifically, the following assumption is made:

Assumption 3 (Independence of the Moderator)

$$(Y_i(t,s), X_i(s)) \perp\!\!\!\perp S_i$$

Under these assumptions, for any $s \in \{0,1\}$:

$$E[Y_i(1,s)|X_i(s)=c] = \lim_{x \downarrow c} E[Y_i(1,s)|X_i(s)=x]$$

$$= \lim_{x \downarrow c} E[Y_i(1,s)|X_i(s)=x, S_i=s] = \lim_{x \downarrow c} E[Y_i|X_i=x, S_i=s]$$

And similarly, again for any $s \in \{0,1\}$:

$$E[Y_i(0,s)|X_i(s)=c] = \lim_{x \uparrow c} E[Y_i|X_i=x, S_i=s]$$

Hence, under these assumptions, the Moderation-in-Discontinuities estimand is identified as such:

$$\begin{aligned} & E[Y_i(1,1) - Y_i(0,1)|X(1)=c] - E[Y_i(1,0) - Y_i(0,0)|X(0)=c] \\ = & \left( \lim_{x \downarrow c} E[Y_i|X_i=x, S_i=1] - \lim_{x \uparrow c} E[Y_i|X_i=x, S_i=1] \right) \\ - & \left( \lim_{x \downarrow c} E[Y_i|X_i=x, S_i=0] - \lim_{x \uparrow c} E[Y_i|X_i=x, S_i=0] \right) \end{aligned} \tag{2}$$

Note that the quantity that identifies the Moderation-in-Discontinuities estimand in equation (2) is precisely the same quantity that identifies the Heterogeneity-in-Discontinuities estimand in equation (1). Hence, the implication is that, given the randomization of the moderator, the Heterogeneity-in-Discontinuities is indeed the Moderation-in-Discontinuities and can be interpreted as such.[4] (Of course the two are not the same in the absence of an exogenous/randomized moderator; in that case, while the Heterogeneity-in-Discontinuities

[4] As described earlier, this also corresponds to the use of the "difference-in-discontinuities" approach to estimate two-period before-after RD effects—i.e. before and after some policy—and assuming time is of no consequence (cites). Such an approach is tantamount to assuming the policy ($S$) is randomly assigned, in which case the Heterogeneity-in-Discontinuities is the Moderation-in-Discontinuities; however, the plausibility of such an assumption in the case of policy change over time is a separate question.

estimand is identified as above, the Moderation-in-Discontinuities estimand must be identified through other means that does not equate with the identifying quantity in equations (1) and (2).)

Given that the quantity that identifies the Heterogeneity-in-Discontinuities estimand also identifies the Moderation-in-Discontinuities estimand when the moderator has been randomized, the estimation procedure can also follow that used in the Heterogeneity-in-Discontinuities context. That is, $E[Y_i|S_i, X_i]$ can again be modeled as follows:

$$\begin{aligned} E[Y_i|S_i, X_i] &= \alpha_0 + \alpha_1\tilde{X}_i + \alpha_2 S_i + \alpha_3\tilde{X}_i \cdot S_i + \\ &\quad \beta_0 T_i + \beta_1 T_i \cdot \tilde{X}_i + \beta_2 T_i \cdot S_i + \beta_3 T_i \cdot \tilde{X}_i \cdot S_i \end{aligned}$$

where $\tilde{X}_i = X_i - c$. Under this model with the additional assumption that the moderator has been randomized, $\beta_2$ now identifies the Moderation-in-Discontinuities estimand of interest. As before, estimation can be done via a local linear regression.

#### 5.2.2. Strategy with Conditionally Independent Moderator

Like the previous, this strategy also follows the standard RD formulation based upon continuous conditional expectation functions. However, in contrast to the previous, this strategy relies upon *conditional* independence of the moderator, along with an accompanying requirement for common support.

Specifically, in addition to Assumption 2, the following assumptions are made:

Assumption 4 (Moderator Conditional Independence)

$$Y_i(t, s) \perp\!\!\!\perp S_i \mid (X_i(s), \boldsymbol{W}_i)$$

*for all* $t \in \{0, 1\}$ *and* $s \in \{0, 1\}$.

Assumption 5 (Moderator Common Support) [5]

$$0 < Pr(S_i = 1|X_i, \boldsymbol{W}_i) < 1$$

Under these assumptions, for any $s \in \{0, 1\}$:

$$E[Y_i(1, s)|X_i(s) = c] = E_W\big[E[Y_i(1, s)|X_i(s) = c, \boldsymbol{W}_i]\big]$$

$$= E_W\big[E[Y_i(1, s)|X_i(s) = c, \boldsymbol{W}_i, S_i = s]\big] = E_W\big[E[Y_i(1, s)|X_i = c, \boldsymbol{W}_i, S_i = s]\big]$$

[5] Note the possibility of weak support, such as $Pr(S_i = 1|X_i, \boldsymbol{W}_i) < 1$, with slightly modified estimand conditional on being (un)moderated.

$$= \lim_{x \downarrow c} E_W \big[ E[Y_i(1,s) | X_i = x, \boldsymbol{W}_i, S_i = s] \big] = \lim_{x \downarrow c} E_W \big[ E[Y_i | X_i = x, \boldsymbol{W}_i, S_i = s] \big]$$

And similarly, again for any $s \in \{0, 1\}$:

$$E[Y_i(0,s) | X_i(s) = c] = \lim_{x \uparrow c} E_W \big[ E[Y_i | S_i = s, X_i = x, \boldsymbol{W}_i] \big]$$

Given that this identification strategy rests upon the moderator conditional independence, estimation requires covariate adjustment. To achieve this, one can build upon the local linear regression framework that is, as already described, commonly used to estimate simple RD effects. Below, we detail several variants of such an estimation approach with varying degrees of flexibility and parameterization.

Local Linear Regression. In specifying a regression model for $E[Y_i | S_i, X_i, \boldsymbol{W}_i]$, a starting point would be a model that allows $S$ and $W$ to each independently (but not interactively) alter the expected relationship between outcome and the running variable. Further, these relationships should also be allowed to vary on each side of the cutoff, following best practices in RD modeling more generally. These criteria imply the following model, which can then be estimated using a local linear regression (e.g. fitting this regression for data within a certain bandwidth of $X$, if using a local linear regression with a rectangular kernel):

$$\begin{aligned} E[Y_i | S_i, X_i, \boldsymbol{W}_i] \;=\; & \alpha_0 + \alpha_1 \tilde{X}_i + \alpha_2 S_i + \alpha_3 \boldsymbol{W}_i + \alpha_4 \tilde{X}_i \cdot S_i + \alpha_5 \tilde{X}_i \cdot \boldsymbol{W}_i + \\ & \beta_0 T_i + \beta_1 T_i \cdot \tilde{X}_i + \beta_2 T_i \cdot S_i + \beta_3 T_i \cdot \boldsymbol{W}_i + \beta_4 T_i \cdot \tilde{X}_i \cdot S_i + \beta_5 T_i \cdot \tilde{X}_i \cdot \boldsymbol{W}_i \end{aligned} \tag{3}$$

where $\tilde{X}_i = X_i - c$. Under this model, $\beta_2$ identifies the Moderation-in-Discontinuities estimand of interest.

As can be seen, even in this baseline scenario only allowing for $S$ and $W$ to have independent influences, and hence ruling out interactions between $S$ and $W$, the estimation specification still requires many other interactions in order to properly line up with the Moderation-in-Discontinuities estimand and its identification. In contrast, prior research attempting to recover estimates of a Moderation-in-Discontinuities effect have implemented RD specifications with the inclusion of select interactions that have not be fully justified or guided by a completely formalized framework (e.g. Bazzi et al., 2020; Desai and Frey, 2021; Olson, 2020).

Local Linear Regression with Regularization. Given that this baseline specification already contains interactions between continuous $X$ and potentially continuous variables contained in $\boldsymbol{W}$, one might be concerned about noise and sensitivity due to overparameterization. Indeed, past research has highlighted the statistical problems induced by

the inclusion of higher-order polynomials of the running variable in a standard RD specification Gelman and Imbens (2019), which suggests taking particular care in the inclusion of interactions between continuous variables in the specification above.

To partially deal with these concerns, estimation of the regression model could also incorporate regularization penalties ($\ell^1$, $\ell^2$, or a combination) on the higher-order terms, with the penalty tuning parameter(s) chosen, for instance, via cross-validation. Taking such an approach, one might then feel more comfortable including even greater flexibility in the regression model specification by including additional interactions and/or polynomial terms. In particular, a natural next step would be to allow $S$ and $W$ to also operate interactively, leading to the following model:

$$\begin{aligned} E[Y_i|S_i, X_i, \boldsymbol{W}_i] &= \alpha_0 + \alpha_1\tilde{X}_i + \alpha_2 S_i + \alpha_3 \boldsymbol{W}_i + \alpha_4 \tilde{X}_i \cdot S_i + \alpha_5 \tilde{X}_i \cdot \boldsymbol{W}_i + \\ &\quad \alpha_6 S_i \cdot \boldsymbol{W}_i + \alpha_7 \tilde{X}_i \cdot S_i \cdot \boldsymbol{W}_i + \\ &\quad \beta_0 T_i + \beta_1 T_i \cdot \tilde{X}_i + \beta_2 T_i \cdot S_i + \beta_3 T_i \cdot \boldsymbol{W}_i + \beta_4 T_i \cdot \tilde{X}_i \cdot S_i + \beta_5 T_i \cdot \tilde{X}_i \cdot \boldsymbol{W}_i + \\ &\quad \beta_6 T_i \cdot S_i \cdot \boldsymbol{W}_i + \beta_7 T_i \cdot \tilde{X}_i \cdot S_i \cdot \boldsymbol{W}_i \end{aligned} \tag{4}$$

In this case, the Moderation-in-Discontinuities estimand of interest is identified by $\beta_2 + \beta_6 E[\boldsymbol{W}_i]$.

Local Linear Regression with a Matched Sample. Choices such as that between model (3) and model (4) above highlight the tradeoff between (a) trying to best approximate the unknown complexities of the true CEFs and (b) flexible functional form mis-specifications leading to estimation problems. The employment of regularization is meant to mitigate such a tradeoff; nonetheless, one may still be worried about functional form restrictions even with regularization. Hence, an alternative would be to undertake a non-parametric approach to conditioning upon the variables contained in $\boldsymbol{W}$ and then proceeding with estimation *after* that covariate adjustment is completed. Specifically, one could apply a local linear regression to matched sample.

The first step for doing this would be, separately on either side of cutpoint, to match units with $S_i = 1$ to units with $S_i = 0$ on $X_i$ and $\boldsymbol{W}_i$. Then, once that is completed, the following model could then be estimated via local linear regression:

$$\begin{aligned} E[Y_i|S_i, X_i] &= \alpha_0 + \alpha_1\tilde{X}_i + \alpha_2 S_i + \alpha_3 \tilde{X}_i \cdot S_i + \\ &\quad \beta_0 T_i + \beta_1 T_i \cdot \tilde{X}_i + \beta_2 T_i \cdot S_i + \beta_3 T_i \cdot \tilde{X}_i \cdot S_i \end{aligned}$$

As can be seen, this is the same model used to estimate the Heterogeneity-in-Discontinuities

estimand. Applied to a matched sample, however, $\beta_2$ identifies the Moderation-in-Discontinuities estimand of interest.

As with the use of matching in any context, one needs to carefully consider the appropriate matching strategy given the data at hand, along with the resulting implications for what is being estimated (i.e. the target estimand to which the matched sample can then pertain). In particular, one could consider either two-way or one-way matching. Two-way or full matching would allow for the target estimand to remain unchanged, though it puts more demands on the data and can prove challenging in the face of imbalanced data and/or asymmetric overlap issues. Hence, one-way matching could be more justifiable or feasible, though the resulting analysis would recover an estimate of a slightly more restricted estimand: either Moderation-in-Discontinuities for the moderated (i.e. for units for which $S = 1$), or Moderation-in-Discontinuities for the unmoderated (i.e. for units for which $S = 0$).

#### 5.2.3. Alternative Strategy with Locally Randomized Running Variable

An alternative to the standard RD formulation based upon continuous conditional expectation functions is a design-based strategy that posits the running variable to be random for a subset of units, typically defined by some interval around the cutoff $c$ (see Cattaneo et al., 2015; Eckles et al., 2020; Li et al., 2015; Mattei and Mealli, 2016). The benefit of this "local randomization" approach, if one feels comfortable with the plausibility of a locally randomized running variable, is the ability to eliminate functional form dependence in the estimation. The Appendix Section A.1 describes this alternative strategy in more detail.

### 5.3. Remarks: Relationship between $S$ and $X$

A key question revolves around whether $S$ affects $X$ and how this affects the inference that can be drawn. First, it is important to note that $S$ should *not* be able to affect $W$, as this would lead to the possibility of conditioning on a variable that is causally posterior to the moderator and hence *prima facie* call into question that credibility of the moderator conditional independence assumption, as in the more general situation of conditioning on a post-treatment variable. However, the identification results above hold regardless of whether or not $S$ has a causal effect on $X$

The most conceptually clear situation would be one where $S$ also does not or cannot affect $X$, which would mean $X_i(0) = X_i(1) = X_i$ and implies a trivial simplification of the results above. This would be the case with an $S$ that is randomized after $X$ is realized (e.g. in an electoral RD, if $X$ is based on previous election margin, randomizing something before

the current election but after the previous election). This could also be the case even if $S$ is not randomized, but there are conceptual/substantive reasons to suspect it does not affect $X$. This would also be the case if proceeding on the basis of a locally randomized running variable.

In the case where $S$ does affect $X$, it is important to further highlight the specific form of the moderator conditional independence assumption, namely that it involves conditioning on $X_i(s)$. This is critical because, whereas $X_i$ is clearly posterior to $S_i$ in this situation, $X_i(s)$ is prior to $S_i$. Further, to the extent that a unit would select into a particular value of $S$, it makes sense that conditioning on $X(s)$ would aid in achieving the exogeneity of $S$, given that $X(s)$ helps to capture the returns from selecting into $S = s$. One might also believe that the nonrandom part of the effect of $S$ on $X$ (i.e. $X(1) - X(0)$) is a function of the variables contained in $\boldsymbol{W}$, in which case conditioning on $\boldsymbol{W}$ along with $X(s)$ also implicitly conditions on $X(s')$.

## 6. CAUTIONS WITH RESPECT TO THE MODERATION-IN-DISCONTINUITIES

As made clear above, the steps and conditions required for identification and estimation of the Moderation-in-Discontinuities are substantially more demanding than that for the Heterogeneity-in-Discontinuities. This section discusses the key hurdles and provides advice to help applied researchers consider whether their context and data allow for a meaningful and plausible investigation of the Moderation-in-Discontinuities.

### 6.1. Conditional Independence

If the moderator has not been randomized, the assumption of conditional independence of the moderator is the core assumption upon which identification of the Moderation-in-Discontinuities rests. Just like with the standard causal identification of an average treatment effect with observational data via selection on observables, conditional independence is a strong and nonrefutable assumption.

Accordingly, researchers should approach this assumption with a sense of healthy skepticism and supplementary analyses to reason through its *degree* of plausibility. That is, unless there exists some administrative or other known mechanism underpinning the moderator, which is unlikely with the common scenarios and moderators in investigations of conditional RD effects, it may be audacious to believe that conditional independence is exactly met. The idea that, in real data, one's control set actually contains every possible required variable is

simply not plausible. This is the same with identification of a simple ATE with observational data. At the same time, however, important evidence and insights can still be generated even in the absence of the full control set. In particular, it is vital to consider (a) the sensitivity of the Moderation-in-Discontinuities estimate to different/additional variables in the control set, which is an empirical matter, and (b) the variables that one believes are important to the theoretical control set but are not available in the data, which is a matter of theory and subject matter expertise. By reasoning through these two dimensions, researchers can offer meaningful insights on the plausibility of and degree to which the moderating variable itself is truly responsible for a difference in RD effects.

In sum, even if one is uncomfortable fully embracing the conditional independence assumption, one can still view the estimation of the Moderation-in-Discontinuities under this approach as a highly principled and structured approach for beginning one's investigation into the causal influence of the moderator. This can be combined with other types of evidence, additional analyses, and careful social scientific reason to provide a clearer (even if imperfect) picture of the causal phenomena potentially at play and to motivate future research with potentially more robust designs.

### 6.2. Common Support

In the standard causal identification of an average treatment effect via selection on observables, the common support (overlap) assumption is equally as important as (and can be viewed as a fundamental continuation of) conditional independence. Unfortunately, however, common support is often not discussed or probed as carefully in applied work—with an exception perhaps being applied work using matching and/or propensity-score-based methods, which naturally lend themselves toward diagnostics focused on overlap.

Common support is critical also for both the identification and estimation of the Moderation-in-Discontinuities, and checking for common support should be a virtually automatic first step for researchers considering an investigation of Moderation-in-Discontinuities. Further, it is important to note that the common support assumption is conditional on both $\boldsymbol{W}$ and $X$. Fortunately, common support is an assumption that can be investigated through a range of diagnostics. In the present case, it is particularly important to check for evidence of common support at the the cutoff, though it is also important to assess overlap across the full range of $X$ within the estimating bandwidth when functional form assumptions are implicit in the estimation strategy.

If there is insufficient common support found with respect to either $X$ or $\boldsymbol{W}$, then it is not

worthwhile to even proceed with estimating a Moderation-in-Discontinuities effect, unless one is comfortable with extrapolation based on extreme reliance on functional form assumptions. In such a case, it is simply not feasible to investigate the Moderation-in-Discontinuities, it is either not identified or otherwise not plausibly estimable with the available data.

### 6.3. Continuity

As noted earlier, the assumption of continuity required for the Heterogeneity-in-Discontinuities estimand is a simple extension of the continuity assumption applied in the standard RD design; the only difference is that the assumption is applied conditional upon observable values of $S$. In contrast, the Moderation-in-Discontinuities estimand requires a more expansive assumption of continuity in the expectation of all potential outcomes $Y(t, s)$, unconditional upon observable values of $S$. That being said, in both cases, the most serious practical threat to continuity remains sorting across the threshold. Furthermore, this threat is due to the possibility that units are sorting across the threshold (i.e. selecting into values of $X$ around the threshold $c$) in order to affect their treatment status $T$. As in the standard RD context, it remains the case for both the Heterogeneity-in-Discontinuities and Moderation-in-Discontinuities that $T$ is defined with respect to the $X$, whereas $S$ is not (and recall that our setup also rules out the possibility of a causal effect of $X$ on $S$).

Hence, the incorporation of the moderator variable should not materially change the nature (or level of threat to) the continuity assumption for either the Heterogeneity-in-Discontinuities or Moderation-in-Discontinuities, relative to the standard RD effect estimand. Indeed, it is hard to think of realistic situations in which the continuity assumption would be met for one of these three estimands (standard RD effect, Heterogeneity-in-Discontinuities, and Moderation-in-Discontinuities) but not met for the others. As such, it makes sense for researchers to approach the assumption of continuity in the same ways they would for the standard RD design, theorizing on the usual concern of sorting across the threshold and applying standard diagnostics (e.g. McCrary, 2008; Cattaneo et al., 2020; Hartman, 2021; for an overview, see Cattaneo and Titiunik, 2021).

## 7. APPLICATIONS

How does existing research stand with respect to the issues discussed in this paper? To address this question, we evaluated previous studies focused on conditional RD effects – both in the electoral setting and outside of it – in light of the framework introduced above.

To start, Table A.1 in Section B.1 summarizes prominent examples of recent research in-

vestigating conditional RD effects across economics, education, and political science. Overall, our review of recent papers suggests the following:

1. Few papers are clear about the specific estimand they are interested in and/or how to interpret their estimates, with respect to the causal role of the conditioning variable.

2. Though some papers include interactions with covariates in their estimation strategy, they do not do so in a formally justified way.

3. Only one of the papers we examined reports any kind of overlap analysis.

4. The control set $\boldsymbol{W}$ is finite and non-exhaustive, meaning that even after our checks, interpretation warrants caution.

We illustrate a more detailed application of our framework using two papers, where we have been able to replicate the research design and extend it with a sufficiently rich set of controls, $\boldsymbol{W}$. We demonstrate how applied research can benefit from our contribution by distinguishing clearly between Heterogeneity-in-Discontinuities and Moderation-in-Discontinuities, and assess the level of confidence with which the conditional RD effects that these applications find can be attributed to a causal effect of the conditional variable (moderator) itself.

### 7.1. Gender Gap in Winning On Persistence

Are women less persistent in running for office when losing an election? A line of very recent work (Bernhard and de Benedictis-Kessner, 2021; Wasserman, 2018, 2021) studies this question in the context of the United States employing a Heterogeneity-in-Discontinuities design. Outside of the US, Cipullo (2021) uses the same design to study the gender gap in persistence in Italian mayoral candidates. In this section, we replicate and extend Cipullo (2021), reporting estimates of the estimands presented earlier.[6]

#### 7.1.1. Original Design and Interpretation

Cipullo (2021), following Wasserman (2018) and Bernhard and de Benedictis-Kessner (2021) in the context of U.S. local elections, studies whether the attrition effect of losing an election

[6]In order to replicate the paper, we collected data on Italian municipal elections ourselves and merged it with the Italian Ministry of Interior's Dataset on Elected Officeholders. Data for most election losers is recorded because most mayoral candidates end up serving as town councillors instead, and are therefore retained in the official data. While both the design and data sources closely follow Cipullo, small discrepancies may remain as we did not have access to the author's replication data.

on the probability to run again is differentially greater for female candidates in U.S. House races and Italian mayoral races. As discussed above, there are multiple reasons why women might suffer from greater attrition – only some of which are directly attributable to gender. Women might, for example, be subject to more negative campaign experiences; they might also, on average, be more averse to competition (Wasserman, 2018). Alternative explanations center on variables correlated with, but distinct from gender: women might be older, run for more or less competitive offices, or attempt to enter politics with less experience. This distinction between possible mechanisms is a textbook case for our framework and underlines the importance of carefully specifying the theorised role of gender, which, in turn, has implications for which estimand is the right quantity of interest.

Though not explicitly spelled out, we consider the implied estimand of the paper to be the Moderation-in-Discontinuities. Cipullo (2021) applies the design in order to examine the 'sticky floor hypothesis', which states that underrepresented groups, such as women, face greater difficulties in being elected; a later interpretation of the key estimate of interest also infers that 'gender differences in future returns from participating in an election depend crucially on the challenges that *women* face [...]' (p. 23)'. Despite being interested in the Moderation-in-Discontinuities, Cipullo (2021) does not discuss the difference between the two estimands. It is worth contrasting this with Wasserman (2018) who, studying the same question in the US, is commendably explicit about the fact that male and female candidates likely differ on a number of dimensions, and therefore intends to estimate the heterogeneity in causal RD effects.[7]

Despite the paper's stated interest in the Moderation-in-Discontinuities, Cipullo (2021)'s key specification intending to capture gender differences in attrition tracks our Heterogeneity-in-Discontinuities estimator closely. The specification uses candidates' margin of victory as the running variable, and assigns the treatment to candidates who barely lost; both treatment and running variable are also interacted with a dummy for female candidates in order to capture the heterogeneity. As a departure from our vanilla estimator, Cipullo (2021) also uses election year fixed effects interacted with both the running variable and the treatment indicator.[8] The proposed specification only tracks the Moderation-in-Discontinuities if con-

[7] Wasserman (2018) proceeds to probe potential mechanisms that depend on covariates potentially correlated with gender, such as women potentially running in more competitive races. She finds that the attrition gap attenuates in elections to more senior offices, thus providing evidence that renders explanations based on fundamental behavioral differences between men and women less likely. This, underlines the interpretation of the gender difference as a Heterogeneity-in-Discontinuities.

[8] Although this specification could be seen as a moderation-in-discontinuities estimator with year indicators as the only variables in the control set, such an interpretation would be far-fetched as it would require that the only difference between men and women is that they run in separate years.

ditional independence holds otherwise, which, based on previously mentioned alternative mechanisms, is unlikely to hold. Overall, the discrepancy between the estimand of interest (MiD) and the estimator (HiD) is illustrative of many papers in the literature.

#### 7.1.2. Applying Our Framework

As a next step, we apply the framework presented above to Cipullo (2021). We assess overlap and fit moderation-in-discontinuities estimators in order to increase our confidence about capturing the intended estimand of interest.

**Introducing Control Set.** We replicate and extend Cipullo's results by collecting our own data on Italian mayors and extending it with a viable control set. Our conditioning variables in this application comprise candidates' age, education level, whether they were born in the municipality or not, and the municipality's logged population. We acknowledge that, in the face of limited data available, this control set is far from exhaustive. Even so, it can serve as a useful first-order check on adjudicating the precise role of gender in studying differential attrition.

**Overlap assessment.** We begin by assessing the overlap of covariates between male and female candidates in our sample. Figure 2 plots the distribution of matched propensity scores for both genders, restricted to matched pairs that fall within a given bandwidth (Appendix C.1 reports the distribution of propensity scores before matching). We observe high overlap and robustness to bandwidth choices. This indicates that male and female candidates in our data set have broadly similiar characteristics, which is desirable, and allows us to proceed with the estimation of Moderation-in-Discontinuities.

**Robustness to MiD Estimators.** Next, we assess whether the gender gap changes in magnitude when we estimate the Moderation-in-Discontinuities (as opposed to the Heterogeneity-in-Discontinuities). Figure 2 reports the estimates at different bandwidths for the vanilla heterogeneity-in-discontinuity specification along with various moderation-in-discontinuity estimators discussed earlier. We see that, overall, the magnitude of the estimate does not change significantly across specifications. In other words, even when we account for the control set, the gender difference in the effect of winning on running again persists.

We emphasize that we should be careful in interpreting the gender gap as a causal effect of gender, because our control set is limited in scope. Nonetheless, applying our framework strengthens confidence in the results and can help to credibly rule out a number of mecha-

nisms that rely on gender correlates (e.g., women being older and therefore less likely to run again). We see this application as a key example of how our framework can help applied researchers distinguish between the two estimands, and how estimation can proceed if the estimand of interest is the Moderation-in-Discontinuities.

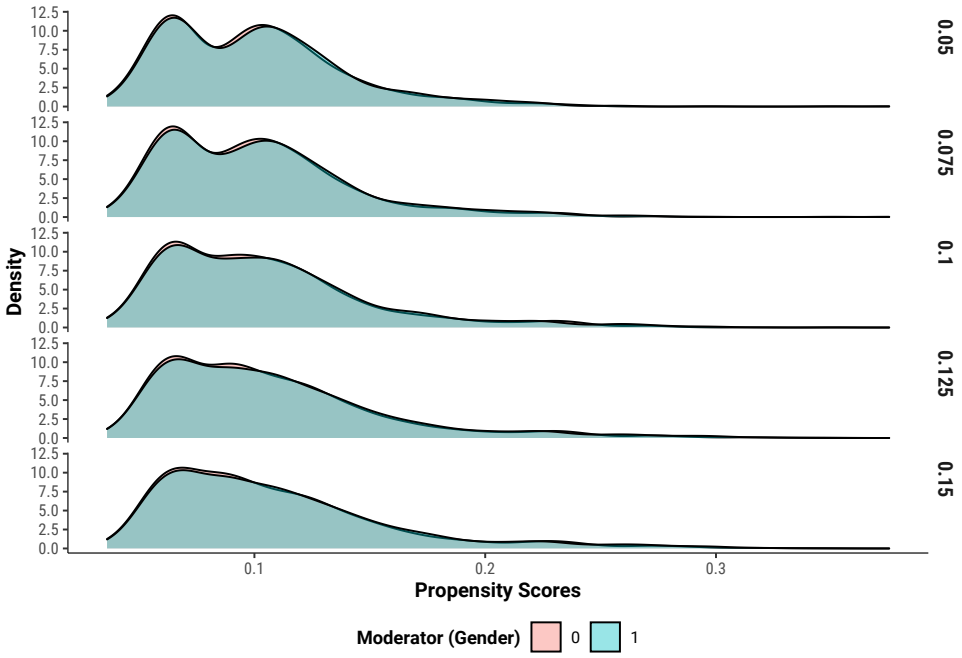


Figure 1: Assessing overlap among observations in Cipullo (2021), by bandwidth. We plot the distribution of propensity scores (measuring how likely an observation is female) after matching on the full control set.

### 7.2. The Effect of Poverty Levels on Right-Wing Party Policies

Our second application studies whether the difference in implemented pro-poor policies between left-wing and right-wing mayors attenuates in poorer towns. Desai and Frey (2021) argue that '[r]ight-wing parties are only competitive in very poor areas if they implement pro-poor policies that voters most often identify with the Left [...]' (p. 2), and offer evidence in the form of a design similar to the Heterogeneity-in-Discontinuities design applied to Brazilian municipalities. We discuss the ambiguity in the intended estimand, and estimate a meaningful difference between the heterogeneity-in-discontinuities and the moderation-in-discontinuities. We believe this application is illustrative of a lack of formalization and theoretical clarity about the estimand that is, unfortunately, common across many papers studying RD effect conditionality.

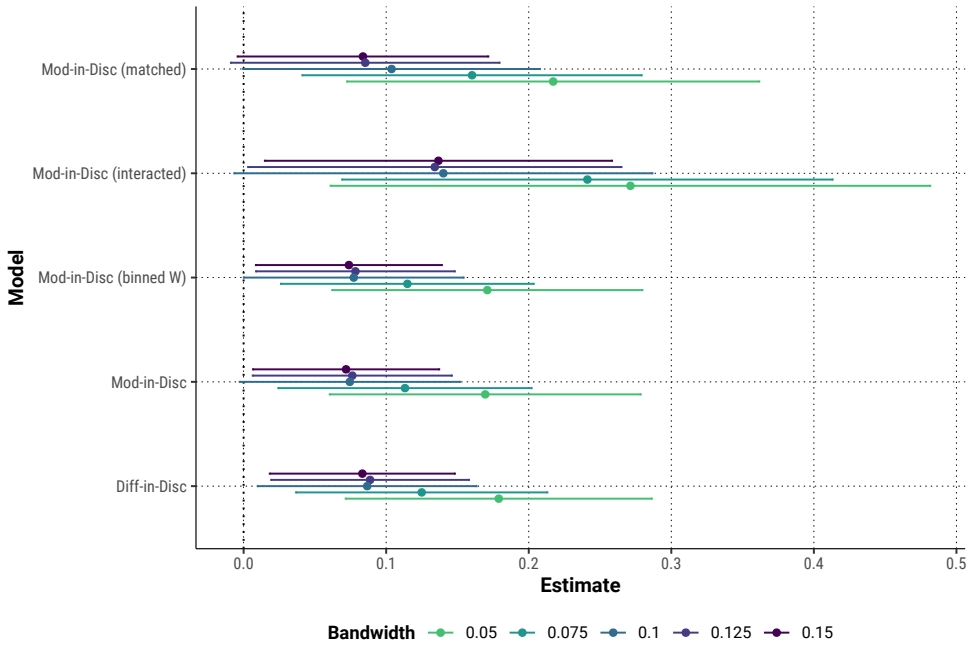


Figure 2: Moderation-in-Discontinuities estimates on the effect of direct primaries on incumbency advantages by gender in Italian Mayoral Elections, following Cipullo (2021)

### 7.2.1. Original Design and Interpretation

Desai and Frey (2021) are interested whether right-wing mayors spend less on pro-poor policies than left-wing mayors, and whether the policy difference attenuates in poorer municipalities. The key theoretical argument posits that right-wing parties need to implement pro-poor policies – defined by municipality spending on education, health, housing, and sanitation – in high-poverty localities in order to match voters' preferences and win. As a result, there is little policy differentiation between left-wing and right-wing winners. But for such a promise to be credible, right-wing parties may also want to field candidates that are more representative of the electorate (e.g., poorer or less educated). In low-poverty towns, by contrast, there is no policy differentiation or education differentiation. The authors hypothesise that the contrast in policy differentiation is *because* of poverty; their empirical hypothesis is informed by a game theoretic model in which the comparative static with respect to poverty drives the key result. But high- and low-poverty towns may differ with respect to other characteristics, such as geography, climate, or literacy. If that is the case, we may observe a policy equivalence in high-poverty areas simply because budgets are more constrained, or because a more remote location requires greater spending on sanitation basics irrespective of being targeted towards poverty, or simply because of a lack of alternative items to spend the budget on.

The paper does not clearly state which estimand it hopes to track empirically. The theoretical model in the paper implies that the quantity of interest is the Moderation-in-Discontinuities – the moderating effect of high-poverty towns on the causal effect of electing a right-wing (vs. left-wing) mayor. On the other hand, when introducing the empirical design, the authors declare their interest in estimating the RD effect of electing a right-wing mayor in the two subsamples (high- and low-poverty), which maps onto the heterogeneity-in-discontinuities estimand. This kind of ambiguity is common across surveyed papers.

A related issue concerning the estimand, following Marshall (2021), is that the design compares election winners of one kind (left-wing) with election winners of another kind (right-wing). If the goal is to estimate the effect of a right-wing mayor independent of their individual characteristics (e.g. age, education etc.), this design would be biased. Instead, we think of the 'treatment' here as a compound treatment of everything that may be different between left-wing and right-wing candidates (following Hall (2015) and Marshall (2021)). For the purpose of this application, we then constrain our discussion about the moderating mechanism to town-level variables, i.e. attributes that differ between high- and low-poverty towns.

The authors use a specification that is similar to, but does not fully map onto our proposed heterogeneity-in-discontinuities estimator. They interact the running variable (margin of victory for right-wing candidate) and treatment indicator (right-wing candidate won) with the conditioning variable, low-poverty; in addition, they also add election year fixed effects and a several contest- and town-level covariates as linear, addive regressors (i.e., without interactions). These additional parameters are not formally justified and leave the implemented estimator without a clear mapping to either main estimand. We also highlight that the authors report results using triangular kernels throughout, and for a limited selection of bandwidths. (We replicated the results using uniform kernels and a wide range of bandwidths in Appendix C.2, and found that the original results are sensitive to higher bandwidths.)

#### 7.2.2. Applying Our Framework

Next, we apply our framework in order to evaluate whether our formal estimators yield divergent results.

**Control Set.** The authors' original data comes with an extensive set of town-level control variables, ranging from municipalities' GDP and geography to past vote shares in higher-order elections. In our replication, we focus on the following municipality-level controls as that are susceptible to feature an independent moderation effect on winners' policy decisions:

the size of the past budget; the vote share the left obtained in past presidential elections; inequality (Gini); GDP per capita; population (logged), and, finally, longitude and latitude.[9] We use this control set to assess overlap and estimate moderation-in-discontinuities effects as introduced earlier in the paper.

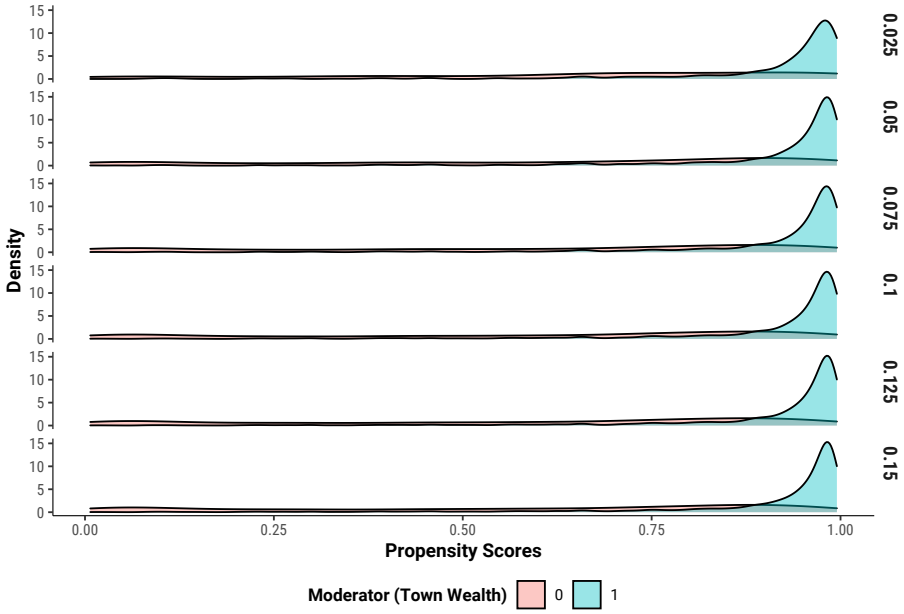


Figure 3: Assessing overlap among observations in Desai and Frey (2021), by bandwidth. We plot the distribution of propensity scores (measuring how likely an observation is in the moderated group) after matching on the full control set (including election year as a continuous covariate).

**Overlap Assessment.** Figure 3 reports the distribution of propensity scores for high-poverty (unmoderated) and low-poverty (moderated) municipalities after matching with replacement (results for matching without replacement and pre-matching are reported in Appendix **??**). Even when allowing for matching with replacement, we do not find good overlap. This result alone threatens the ability to interpret RD effect conditionality as the causal effect of the moderator, as it suggests that wealthy and poor municipalities differ significantly along the dimensions captured by our control set. The example highlights the importance of the overlap check to assess the plausibility of alternative moderating variables that may account for the contrast.

[9] We also leave out year fixed effects because dealing with time as a conditioning variable is difficult. Appendix C.3 shows results where we retain longitude, latitude and population as the only conditioning variables, with similar results to Figure 4.

**Robustness to MiD Estimators.** Despite finding little overlap in the earlier check, we proceed with fitting our estimators for the purpose of illustrating our framework. We caution, however, that applied researchers ought not to interpret MiD estimates as credible if the overlap is as sparse as in this case. Figure 4 presents our estimates. Our heterogeneity-in-differences estimate matches the original paper's results close to the authors' preferred bandwidths (0.052), though our confidence intervals are larger due to the omission of linearly added control variables and year fixed effects, as well as the use of a uniform kernel. In wealthier towns, electing a right-wing mayor (as opposed to electing a left-wing mayor) decreases the share of pro-poor spending in the local budget by a greater magnitude. As we move towards wider bandwidths, however, the heterogeneity-in-discontinuities estimates attenuate towards zero.

In the vanilla MiD, the sign of the coefficient of interest flips compared to previous results: the effect of a right-wing mayor on pro-poor spending *grows* in richer towns. Across all bandwidths, however, these estimates are statistically insignificant. When matching without replacement, we get results that are substantively similar to the original Heterogeneity-in-Discontinuities estimand; this should not be surprising given the poor overlap. Finally, when matching with replacement, the estimates hover around zero for most bandwidths; for the very small bandwidth of 0.025, the moderation effect grows very large in magnitude and positive.

Our analysis points to the following conclusions. First, the application highlights the importance of carefully and formally distinguishing between the two estimands of interest. Second, we stress the importance of assessing observations' overlap with respect to the moderator in order to justify any credible interpretation of estimates as moderation-in-discontinuities. Third, even aside from overlap concerns, the moderation-in-discontinuities estimates may be fundamentally different from the heterogeneity-in-discontinuities ones, casting doubt on clear causal interpretations of the moderation 'effect'. In this particular application, the results suggest we cannot rule out the possibility that a factor correlated with town wealth (e.g., geography) may be responsible for the observed heterogeneity.

## 8. DISCUSSION AND CONCLUSIONS

In this paper, we introduce and discuss an important distinction between two quantities of interest when assessing effect conditionality in regression discontinuity designs. We distinguish between the Heterogeneity-in-Discontinuities estimand, which maps onto the local average treatment effect conditional on a given characteristic (e.g, the effect of winning an election

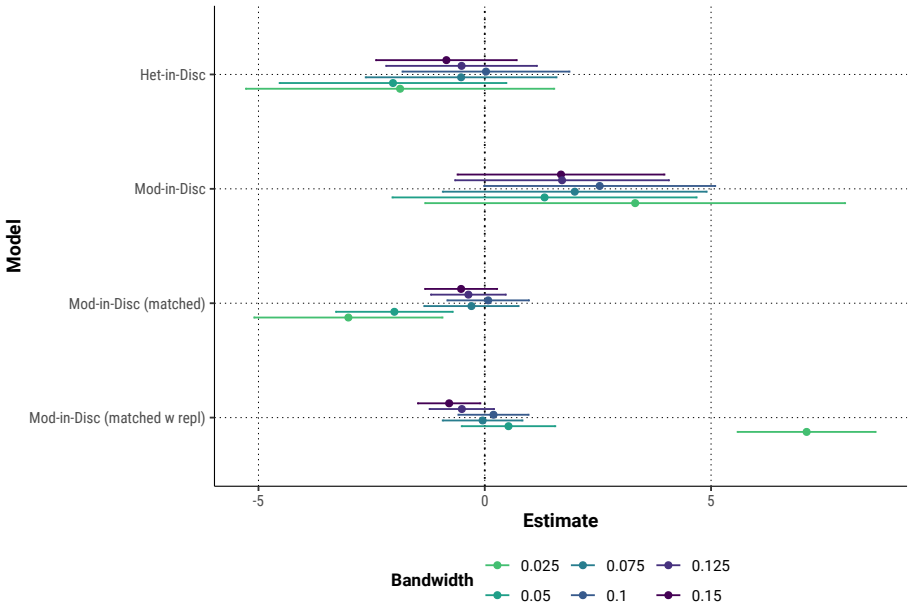


Figure 4: Moderation-in-Discontinuities estimates on the moderation effect of town poverty on the effect of electing a right-wing mayor on pro-poor policies, following Desai and Frey (2021).

for women in the sample), and the Moderation-in-Discontinuities, which recovers the moderation effect of the defining characteristic that causes the treatment effect to change (the effect of winning an election for women *because* they are women). Separating the two estimands is not just important for conceptual reasons, but can help researchers and policymakers draw more precise conclusions from their findings.

We introduce a formalized framework to describe the two estimands in greater detail and offer multiple strategies to recover the Moderation-in-Discontinuities under conditional independence assumptions. In light of existing applied research often being ambiguous about the estimand of interest, we offer practical advice and apply our framework to two recent papers. Our results illustrate the need to formalize and discuss the estimand of interest when targeting RD effect conditionality; when interested in Moderation-in-Discontinuities, researchers should also heed the importance of checking overlap assumptions and the use of appropriate estimators that track the estimand.

Our work comes with an important caveat. We stress that in many settings, the conditional independence assumption may not be credible, or gathering an appropriate control set may not be feasible. At other times, our Moderation-in-Discontinuities estimators may be too demanding in terms of statistical power. In those cases, a credible estimate of the Moderation-in-Discontinuities may not be possible, and researchers should discuss this pos-

sibility openly rather than returning to the Heterogeneity-in-Discontinuities. Despite these challenges, our frameework introduces new methodological vocabulary, an important conceptual distinction along with useful estimation strategies that can benefit researchers across many social science settings where RD effect conditionality is of interest.

## References


Abou-Chadi, T. and Krause, W. (2020). The causal effect of radical right success on mainstream parties' policy positions: A regression discontinuity approach. *British Journal of Political Science*, 50(3):829–847.

Anagol, S. and Fujiwara, T. (2016). The Runner-Up Effect. *Journal of Political Economy*, 124(4):927–991.

Anderson, M. L. (2014). Subways, Strikes, and Slowdowns: The Impacts of Public Transit on Traffic Congestion. *The American Economic Review*, 104(9):2763–2796.

Bansak, K. (2021). Estimating causal moderation effects with randomized treatments and non-randomized moderators. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 184(1):65–86.

Barrow, L., Sartain, L., and de la Torre, M. (2020). Increasing Access to Selective High Schools through Place-Based Affirmative Action: Unintended Consequences. *American Economic Journal: Applied Economics*, 12(4):135–163.

Bazzi, S., Koehler-Derrick, G., and Marx, B. (2020). The institutional foundations of religious politics: Evidence from indonesia. *The Quarterly Journal of Economics*, 135(2):845–911.

Becker, S. O., Egger, P. H., and Von Ehrlich, M. (2013). Absorptive capacity and the growth and investment effects of regional transfers: A regression discontinuity design with heterogeneous treatment effects. *American Economic Journal: Economic Policy*, 5(4):29–77.

Bernhard, R. and de Benedictis-Kessner, J. (2021). Men and women candidates are similarly persistent after losing elections. *Proceedings of the National Academy of Sciences*, 118(26).

Bohlken, A. T. (2018). Targeting Ordinary Voters or Political Elites? Why Pork Is Distributed Along Partisan Lines in India. *American Journal of Political Science*, 62(4):796–812.

Bronzini, R. and Iachini, E. (2014). Are incentives for r&d effective? evidence from a regression discontinuity approach. *American Economic Journal: Economic Policy*, 6(4):100–134.

Brown, R., Mansour, H., O'Connell, S., and Reeves, J. (2020). Gender differences in political career progression. *IZA Document Paper No. 12569*.

Calonico, S., Cattaneo, M. D., Farrell, M. H., and Titiunik, R. (2019). Regression discontinuity designs using covariates. *Review of Economics and Statistics*, 101(3):442–451.

Card, D. and Giuliano, L. (2016). Can Tracking Raise the Test Scores of High-Ability Minority Students? *American Economic Review*, 106(10):2783–2816.

Cattaneo, M. D., Frandsen, B. R., and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the us senate. *Journal of Causal Inference*, 3(1):1–24.

Cattaneo, M. D., Jansson, M., and Ma, X. (2020). Simple local polynomial density estimators. *Journal of the American Statistical Association*, 115(531):1449–1455.

Cattaneo, M. D. and Titiunik, R. (2021). Regression discontinuity designs. *arXiv preprint arXiv:2108.09400*.

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., and Morris, R. E. (2018). The internal and external validity of the regression discontinuity design: A meta-analysis of 15 within-study comparisons. *Journal of Policy Analysis and Management*, 37(2):403–429.

Chicoine, L. E. (2017). Homicides in Mexico and the expiration of the US federal assault weapons ban: A difference-in-discontinuities approach. *Journal of economic geography*, 17(4):825–856.

Cipullo, D. (2021). Gender Gaps in Political Careers: Evidence from Competitive Elections.

Cook, T. D. and Wong, V. C. (2008). Empirical tests of the validity of the regression discontinuity design. *Annales d'Economie et de Statistique*, pages 127–150.

De Benedetto, M. A. and De Paola, M. (2019). Term limit extension and electoral participation. Evidence from a diff-in-discontinuities design at the local level in Italy. *European Journal of Political Economy*, 59:196–211.

de Benedictis-Kessner, J. (2018). Off-cycle and out of office: Election timing and the incumbency advantage. *The Journal of Politics*, 80(1):119–132.

Desai, Z. and Frey, A. (2021). Can Descriptive Representation Help the Right Win Votes from the Poor? Evidence from Brazil. *American Journal of Political Science*, n/a(n/a).

Dunaiski, M. (2021). Is compulsory voting habit-forming? Regression discontinuity evidence from Brazil. *Electoral Studies*, 71:102334.

Dunning, T. (2012). *Natural experiments in the social sciences: a design-based approach*. Cambridge University Press.

Eckles, D., Ignatiadis, N., Wager, S., and Wu, H. (2020). Noise-induced randomization in regression discontinuity designs. *arXiv preprint arXiv:2004.09458*.

Eggers, A. C. (2015). Proportionality and turnout: Evidence from French municipalities. *Comparative Political Studies*, 48(2):135–167.

Eggers, A. C., Fowler, A., Hainmueller, J., Hall, A. B., and Snyder Jr, J. M. (2015). On the validity of the regression discontinuity design for estimating electoral effects: New evidence from over 40,000 close races. *American Journal of Political Science*, 59(1):259–274.

Eggers, A. C. and Spirling, A. (2017). Incumbency effects and the strength of party preferences: Evidence from multiparty elections in the united kingdom. *The Journal of Politics*, 79(3):903–920.

Feigenbaum, J. J., Fouirnaies, A., Hall, A. B., et al. (2017). The majority-party disadvantage: revising theories of legislative organization. *Quarterly Journal of Political Science*, 12(3):269–300.

Frölich, M. and Huber, M. (2019). Including covariates in the regression discontinuity design. *Journal of Business & Economic Statistics*, 37(4):736–748.

Fujiwara, T. (2011). A Regression Discontinuity Test of Strategic Voting and Duverger's Law. *Quarterly Journal of Political Science*, 6(3-4):197–233.

Gelman, A. and Imbens, G. (2019). Why high-order polynomials should not be used in regression discontinuity designs. *Journal of Business & Economic Statistics*, 37(3):447–456.

Gerber, A. S. and Green, D. P. (2012). *Field Experiments: Design, Analysis, and Interpretation*. New York: W. W. Norton & Company.

Gerber, A. S., Kessler, D. P., and Meredith, M. (2011). The Persuasive Effects of Direct Mail: A Regression Discontinuity Based Approach. *The Journal of Politics*, 73(1):140–155.

Grembi, V., Nannicini, T., and Troiano, U. (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics*, pages 1–30.

Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, 69(1):201–209.

Hansen, B. (2015). Punishment and Deterrence: Evidence from Drunk Driving. *The American Economic Review*, 105(4):1581–1617.

Hartman, E. (2021). Equivalence testing for regression discontinuity designs. *Political Analysis*, 29(4):505–521.

Hsu, Y.-C. and Shen, S. (2019). Testing treatment effect heterogeneity in regression discontinuity designs. *Journal of Econometrics*, 208(2):468–486.

Imai, K. and Ratkovic, M. (2013). Estimating treatment effect heterogeneity in randomized program evaluation. *The Annals of Applied Statistics*, 7:443–470.

Imbens, G. W. and Lemieux, T. (2008). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, 142(2):615–635.

Jenkins, J. M., Farkas, G., Duncan, G. J., Burchinal, M., and Vandell, D. L. (2016). Head start at ages 3 and 4 versus head start followed by state pre-k: Which is more effective? *Educational evaluation and policy analysis*, 38(1):88–112.

Kantorowicz, J. and Köppl–Turyna, M. (2019). Disentangling the fiscal effects of local constitutions. *Journal of Economic Behavior & Organization*, 163:63–87.

Kaplan, E., Saltiel, F., and Urzúa, S. S. (2019). Voting for Democracy: Chile's Plebiscito and the Electoral Participation of a Generation. Technical report, National Bureau of Economic Research.

Köppl-Turyna, M. and Kantorowicz, J. (2020). The effect of quotas on female representation in local politics. Technical report, Research Paper.

Lalive, R. (2007). Unemployment benefits, unemployment duration, and post-unemployment jobs: A regression discontinuity approach. *American Economic Review*, 97(2):108–112.

Li, F., Mattei, A., Mealli, F., et al. (2015). Evaluating the causal effect of university grants on student dropout: evidence from a regression discontinuity design using principal stratification. *Annals of Applied Statistics*, 9(4):1906–1931.

Lopes da Fonseca, M. (2017). Identifying the source of incumbency advantage through a constitutional reform. *American Journal of Political Science*, 61(3):657–670.

Marshall, J. (2019). When can close election RDDs identify the effects of winning politician characteristics? *Working Paper*.

Mattei, A. and Mealli, F. (2016). Regression discontinuity designs as local randomized experiments. *Observational Studies*, 2:156–173.

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

Micozzi, J. P. and Lucardi, A. (2021). How valuable is a legislative seat? incumbency effects in the argentine chamber of deputies. *Political Science Research and Methods*, 9(2):414–429.

Neyman, J. (1923). Sur les applications de la théorie des probabilités aux experiences agricoles: Essai des principes. *Roczniki Nauk Rolniczych*, 10:1–51.

Núñez, L. (2018). Do clientelistic machines affect electoral outcomes? mayoral incumbency as a proxy for machine prowess. *Electoral Studies*, 55:109–119.

Olson, M. P. (2020). The direct primary and the incumbency advantage in the us house of representatives. *Quarterly Journal of Political Science*, 15(4):483–506.

Pettersson-Lidbom, P. (2008). Do Parties Matter for Economic Outcomes? A Regression-Discontinuity Approach. *Journal of the European Economic Association*, 6(5):1037–1056.

Pop-Eleches, C. and Urquiola, M. (2013). Going to a better school: Effects and behavioral responses. *American Economic Review*, 103(4):1289–1324.

Ratkovic, M. (2021). Subgroup analysis: Pitfalls, promise, and honesty. In Druckman, J. N. and Green, D. P., editors, *Advances in Experimental Political Science*, chapter 15, pages 271–288. Cambridge University Press.

Ratkovic, M. and Tingley, D. (2017). Sparse estimation and uncertainty with application to subgroup analysis. *Political Analysis*, 25(1):1–40.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688.

Rubin, D. B. (1980). Comment: Randomization analysis of experimental data: The Fisher randomization test. *Journal of the American Statistical Association*, 75(371):591–593.

Sells, C. J. (2020). Building parties from city hall: Party membership and municipal government in brazil. *The Journal of Politics*, 82(4):1576–1589.

Thistlethwaite, D. L. and Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational psychology*, 51(6):309.

Titiunik, R. (2021). Natural experiments. In Druckman, J. N. and Green, D. P., editors, *Advances in Experimental Political Science*, chapter 6, pages 103–129. Cambridge University Press.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Wasserman, M. (2018). Gender differences in politician persistence. *Available at SSRN 3370587*.

Wasserman, M. (2021). Up the political ladder: Gender parity in the effects of electoral defeats. In *AEA Papers and Proceedings*, volume 111, pages 169–73.

Yoder, J., Handan-Nader, C., Myers, A., Nowacki, T., Thompson, D. M., Wu, J. A., Yorgason, C., and Hall, A. B. (2021). How Did Absentee Voting Affect the 2020 US Election?

# A. ADDITIONAL METHODS

## A.1. Locally Randomized Running Variable and Conditionally Independent Moderator

An alternative to the standard RD formulation based upon continuous conditional expectation functions is a design-based strategy that posits the running variable to be random for a subset of units, typically defined by some interval around the cutoff $c$ (see Cattaneo et al., 2015; Eckles et al., 2020; Li et al., 2015; Mattei and Mealli, 2016). The benefit of this "local randomization" approach, if one feels comfortable with the plausibility of a locally randomized running variable, is the ability to eliminate functional form dependence in the estimation. Applying this approach to the Moderation-in-Discontinuities would proceed with the following local randomization assumption, adapted from Mattei and Mealli (2016):

Assumption 6 (Local Randomization of Running Variable) [10]

*For all $i \in \mathcal{U}_c$, where $\mathcal{U}_c$ denotes a subset of units,*

$$Pr(X_i|Y_i(t,s), S_i, \boldsymbol{W}_i) = Pr(X_i)$$

*for all $t \in \{0,1\}$ and $s \in \{0,1\}$.*

Note that as a part of this assumption, it is implied that among the units belonging to $\mathcal{U}_c$, there is no relationship between $S$ and $X$, and hence $X_i(0) = X_i(1) = X_i$. In addition to local randomization of the running variable, the following assumptions are also necessary:

Assumption 7 (Local Moderator Conditional Independence)

$$Y_i(t,s) \perp\!\!\!\perp S_i \mid \boldsymbol{W}_i$$

*for all $i \in \mathcal{U}_c$ and for $t \in \{0,1\}$ and $s \in \{0,1\}$.*

Assumption 8 (Local Moderator Common Support)

$$0 < Pr(S_i = 1|\boldsymbol{W}_i) < 1$$

*for all $i \in \mathcal{U}_c$.*

[10] In addition, the local randomization approach to regression discontinuities requires two additional enabling assumptions. The first (which Mattei and Mealli (2016) refer to as "local overlap") is the existence of the subset $\mathcal{U}_c$ defined such that for each $i \in \mathcal{U}_c$, $Pr(X_i \leq c) > \epsilon$ and $Pr(X_i > c) > \epsilon$ for some sufficiently large $\epsilon > 0$. This assumption implies that each unit in the subset has a non-zero probability of assignment to either of the treatment conditions. The second additional assumption is a modification of the classic Stable Unit Treatment Value Assumption (SUTVA) attributable to Rubin (1980). This modified assumption (which Mattei and Mealli (2016) refer to as "local RD-SUTVA") states that for each $i \in \mathcal{U}_c$, consider two treatment statuses $T_i' = \mathbf{1}(X_i' \leq c)$ and $T_i'' = \mathbf{1}(X_i'' \leq c)$, with possibly $X_i' \neq X_i''$; if $T_i' = T_i''$, then $Y_i(\boldsymbol{X}') = Y_i(\boldsymbol{X}'')$, where $Y_i(\boldsymbol{X})$ refers to potential outcomes defined as a function of the vector of running variable values $\boldsymbol{X}$ for the full subset. This assumption implies that there is no interference between units and that potential outcomes depend upon the running variable solely through the treatment status, and hence allows for $Y_i(\boldsymbol{X})$ to be simplified as $Y_i(t)$ for each unit $i \in \mathcal{U}_c$. Different variants of these sets of assumptions have also been presented in other related work (Cattaneo et al., 2015; Eckles et al., 2020; Li et al., 2015).

Note that Assumption 6 combined with Assumption 8 implies a local version of $0 < Pr(S_i = 1|X_i, \boldsymbol{W}_i) < 1$, i.e. common support conditional on both $\boldsymbol{W}$ and $X$, as before.

As mentioned above, the benefit of proceeding based on the assumption of a locally randomized running variable, if it can be deemed plausible, is the ability to eliminate functional form dependence in the estimation. Specifically, the Moderation-in-Discontinuities need not be estimated on the basis of paramerizing the expected value of $Y$ conditional on $S$, $X$, and $\boldsymbol{W}$. Instead, the amount of parameterization can be limited, even allowing for entirely nonparametric estimation.

Specifically, one can proceed with the standard approach where $\mathcal{U}_c$ is defined as the subset of units whose running variable values fall within some (symmetric) interval around the cutoff $c$. For this subset, one can then adapt the nonparametric framework presented by Bansak (2021) for estimating causal moderation effects given randomization of a treatment and non-randomization of a moderator. In the present context, this would involve first splitting the data into two subsets—one in which $i \in \mathcal{U}_c$ and $X_i \leq c$, and one in which $i \in \mathcal{U}_c$ and $X_i > c$—and then estimating separately for each of those subsets the causal effect of $S$ on $Y$ via some covariate adjustment strategy conditioning on $\boldsymbol{W}$, which could include nonparametric methods like matching. The difference between these two within-subset estimates would then comprise the Moderation-in-Discontinuities estimate (see Bansak, 2021, for more details).

# B. EXISTING LITERATURE

## B.1. Literature Review

Table 1 summarizes recent studies using Heterogeneity-in-Discontinuities designs.

| Authors | Journal | Treatment | Outcome | Heterogeneity Set | Notes |
|---|---|---|---|---|---|
| Pop-Eleches and Urquiola (2013) | AER | Better School | Student Test Scores | Intial School Performance | |
| Bronzini and Iachini (2014) | AEJ:EP | Subsidies | Investment | Firm Size | |
| Card and Giuliano (2016) | AER | High-Performance School | Student Test Scores | Minority Status | |
| Grembi et al. (2016) | AEJ:Applied | Fiscal Rules | Fiscal Outcomes | Time | (diff. framew |
| Eggers and Spirling (2017) | JOP | Being Elected | Winning Again | Party Competition | |
| de Benedictis-Kessner (2018) | JOP | Being Elected | Winning Again | On-Off-Cycle | |
| Bazzi et al. (2020) | QJE | Land Expropriation | Islamist Strength | Expropriation Intensity | |
| Bohlken (2018) | AJPS | Being Elected | Project Expenditure | Governing Party | |
| Micozzi and Lucardi (2021) | PSRM | Being Elected | Future Career Outcomes | Party Type | |
| Olson (2020) | Being Elected | Winning Again | Nomination Process | | |
| Barrow et al. (2020) | AEJ:Applied | Selective School | Test Scores | Socioeconomic status | |
| Sells (2020) | JOP | Being Elected | Party Membership | Party Type | |
| Wasserman (2018) | ReEconStat | Being Elected | Running Again | Gender | |
| Wasserman (2021) | AEA Proceedings | Being Elected | Running Again | Gender | |
| Bernhard and de Benedictis-Kessner (2021) | PNAS | Being Elected | Running Again | Gender | |
| Desai and Frey (2021) | AJPS | Right-Wing Elected | Pro-Poor Spending | Town Wealth | |
| Brown et al. (2020) | WP | Being Elected (State) | Being Elected (Congress) | Gender | |
| Cipullo (2021) | WP | Being Elected | Running Again | Gender | |
| McCrain et al (2021) | WP | Being Elected (State) | Being Elected (Congress) | Professionalisation | |

Table 1: Summary of recent papers using heterogeneity-in-discontinuity designs

# C. ADDITIONAL ROBUSTNESS CHECKS

## C.1. Propensity Score Distributions Before Matching

In this Appendix, we report the distribution of propensity scores *before* any matching for both of our applications. The results are consistent with the general conclusions from the two applications: in the case of Cipullo (2021), the overlap is already pretty good before any matching, whereas in the case of Desai and Frey (2021), wealthy and poor municipalities offer a stark difference. We also report the overlap for the case of Desai and Frey (2021) after matching without replacement. Unsurprisingly, this method does not improve overlap very much, since matching without replacement does not allow us to find high-quality matches for most observations.

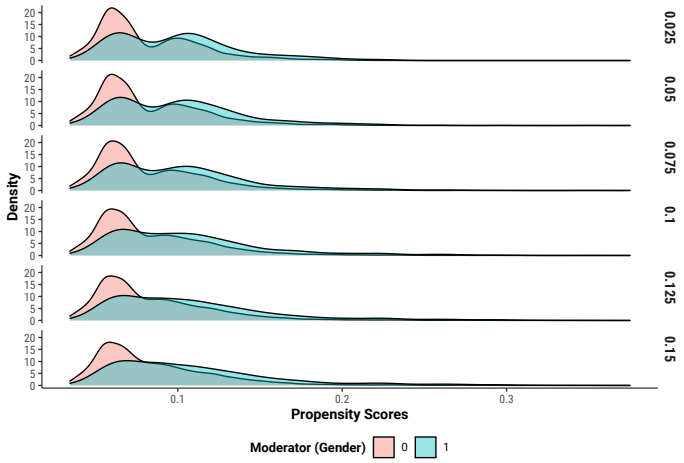


Figure 5: Assessing overlap among observations before matching in Desai and Frey (2021) using propensity scores (measuring how likely an observation is in the moderated group)

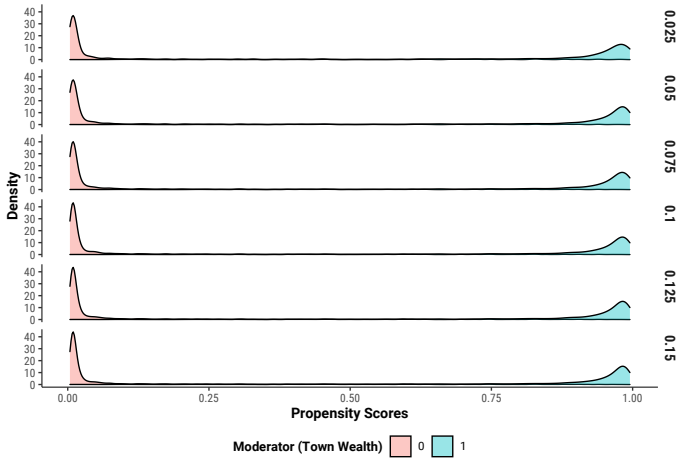


Figure 6: Assessing overlap among observations before matching in Desai and Frey (2021) using propensity scores (measuring how likely an observation is in the moderated group)

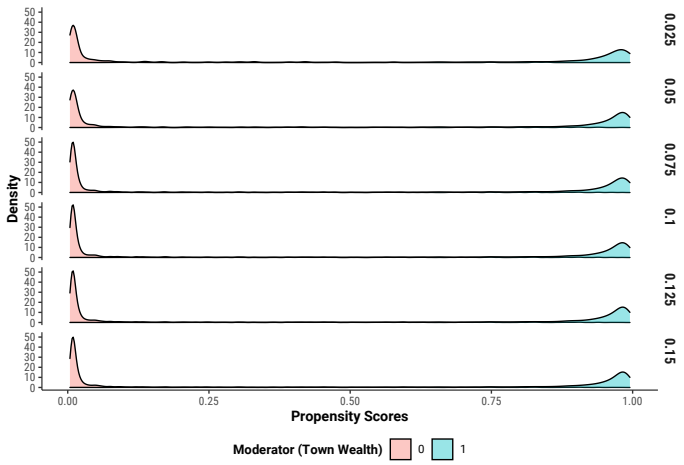


Figure 7: Assessing overlap among *matched* observations (without replacement) in Desai and Frey (2021) using propensity scores (measuring how likely an observation is in the moderated group)

## C.2. Desai and Frey (2021): Sensitivity to Kernel Checks and Bandwidth

Throughout our applications, we use uniform kernels around the threshold in estimating the Heterogeneity-in-Discontinuities and Moderation-in-Discontinuities. In the original paper, Desai and Frey (2021) use triangular kernels as default estimation procedure. In Figure **??**, we report our estimates from the vanilla Heterogeneity-in-Discontinuities estimation specification (with covariates and year fixed effects) for both kernel estimators across a wide range of the bandwidth.

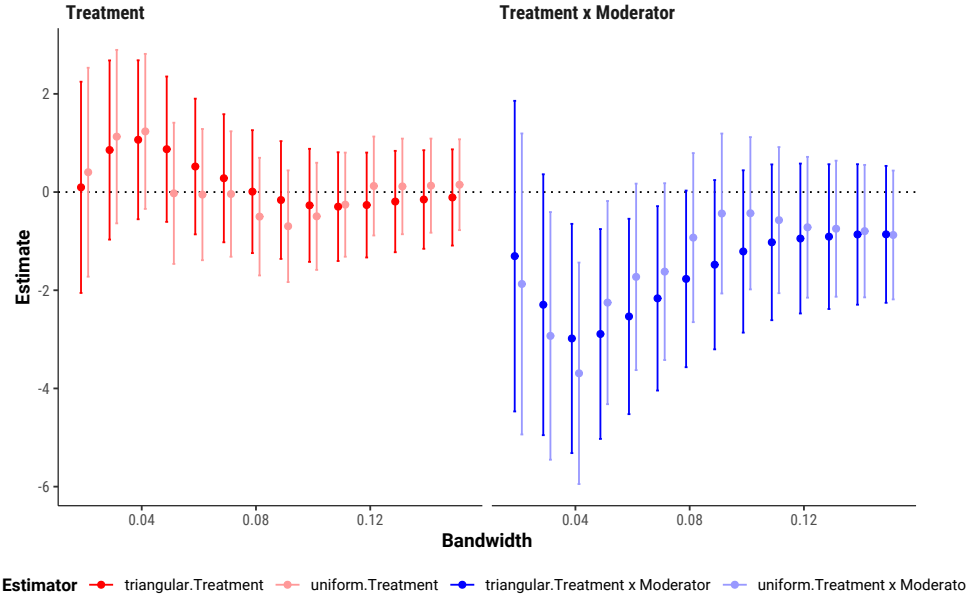


Figure 8: Assessing overlap among matched observations (without replacement) in Desai and Frey (2021) using propensity scores (measuring how likely an observation is in the moderated group)

## C.3. Desai and Frey (2021): Robustness to Smaller Control Set

Finally, we also report our results from the Moderation-in-Discontinuities estimators when using a control set that is limited to longitude, latitude, and population. We observe results that are very similar to our main results in the paper, highlighting the role that these three variables play in distinguishing between the Heterogeneity-in-Discontinuities and the Moderation-in-Discontinuities.

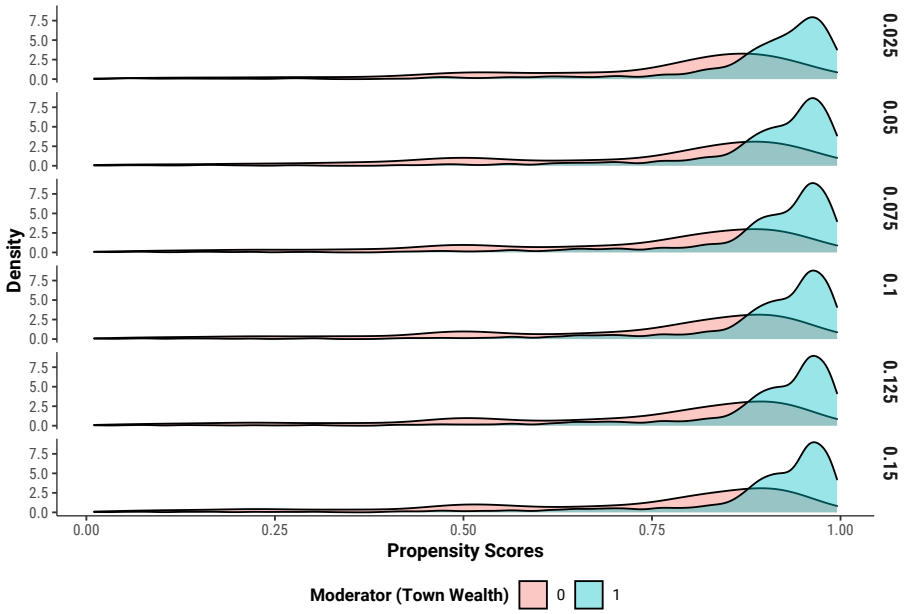


Figure 9: Assessing overlap among matched observations (with replacement) in Desai and Frey (2021) using propensity scores (measuring how likely an observation is in the moderated group) and a limited control set

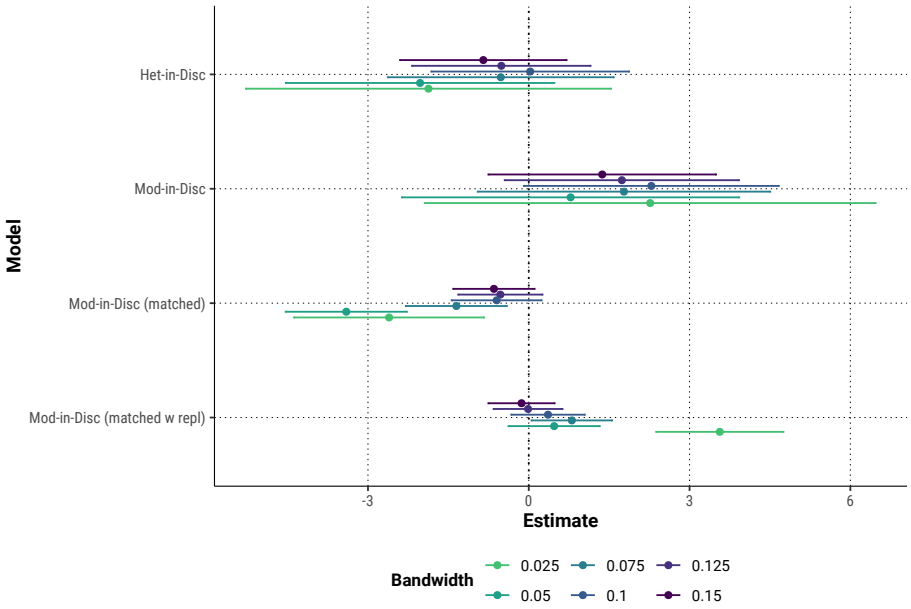


Figure 10: Moderation-in-Discontinuities estimates on the moderation effect of town poverty on the effect of electing a right-wing mayor on pro-poor policies, following Desai and Frey (2021) and using a limited control set.