Effect Heterogeneity and Causal Attribution in Regression Discontinuity Designs [*]

Kirk Bansak[†] Tobias Nowacki[‡]

Abstract

Research investigating subgroup differences in treatment effects recovered using regression discontinuity (RD) designs has become increasingly popular. For instance, scholars have investigated 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 subgroup differences in treatment effects as a causal result of the moderating characteristic? In this study, we explore the difference between RD effect conditionality that is simply associated with versus causally driven by another variable. To make this distinction explicit and formal, we define two alternative estimands and lay out identification assumptions required for each, along with corresponding estimation procedures. In doing so, we highlight how investigating RD effect conditionality that is causally driven by another variable involves several additional challenges related to interpretation, identification, and estimation. We apply our framework to recent studies and offer practical advice for applied researchers considering these alternative quantities of interest.

[*] Acknowledgments: We thank Max Goplerud, Cyrus Samii, Dan Thompson and participants at the 2021 Stanford Causal Science Center Conference and 2022 MPSA Conference for helpful feedback. We thank Gemma Dipoppa for generously sharing her data on Italian mayoral elections with us.

[†] Assistant Professor, Department of Political Science, University of California, Berkeley, 210 Social Sciences Building, Berkeley, CA 94720-1950, United States. E-mail: kbansak@berkeley.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 or receipt of a 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 (e.g., age, vote shares, population sizes). Under certain conditions, a causal effect of the treatment can then be reliably estimated by comparing the trends of units close to and on either side of the threshold. 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).

As usage of the RD design has become more pervasive across the social sciences, with researchers often iterating in greater 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 expanding focus in experimental research on conditional average treatment effects and treatment effect heterogeneity (see e.g. Gerber and Green, 2012; Imai and Ratkovic, 2013; Ratkovic, 2021; Ratkovic and Tingley, 2017; Wager and Athey, 2018).

Along these lines, a number of electoral RD studies in political science have investigated

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

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% more for women than it does for men, although this gender gap may attenuate in more senior electoral settings (Bernhard and de Benedictis-Kessner, 2021; Wasserman, 2021).

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? This is a distinction that is important from both policy and theoretical perspectives. For instance, the causal source of the incumbency effect gender gap has implications for how widely it would manifest across different contexts – a point we return to in the next section.

Yet 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, 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. Elsewhere, the motivation for investigating a particular conditioning variable is connected to a theorized mechanism that strongly implies a causal influence of that moderator variable, but the empirical analyses are undertaken in such a way that suggests otherwise. 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 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 study seeks to help address these issues, providing theory and methods to allow researchers 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 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* (HiD), refers to the difference in RD effects across observed levels of some other variable. The second, which we term *Moderation-in-Discontinuities* (MiD), refers to the difference in RD effects that is actually caused by (in a counterfactual sense) the other variable. We lay out different sets of identification assumptions that are required for each estimand, along with corresponding estimation procedures.

Along the way, we highlight several additional challenges related to the MiD, involving interpretation, identification, and estimation. Accordingly, we emphasize the caution that applied researchers must take in investigating evidence of RD effect conditionality, and highlight the importance of careful consideration of one's data, research design, and quantity of interest. We discuss these issues and offer 'best practice' advice for applied researchers, including diagnostics for considering the plausibility of the identification assumptions behind the MiD. To illustrate our framework and further highlight the challenges involved, we present multiple applications with varying degrees of plausibility that the MiD is identified. We conclude that it may be possible to generate reasonable evidence of the MiD in some contexts, but that in other contexts there is little to no plausibility of doing so, and applied researchers must be very careful in their analysis and interpretation of RD effect conditionality.

2 MOTIVATION AND RELATED WORK

2.1 Motivating Examples

Both the HiD and the MiD 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. 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 an election on candidates' decision to run again (Bernhard and de Benedictis-Kessner, 2021; Cipullo, 2021; Wasserman, 2018). A gender difference in candidate persistence might contribute to the underrepresentation of women in politics. Various 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 persistence gap. Equally plausibly, however, women might be less likely to run again if, on average, they are older, or have more 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 essential for policy implications: if being a women has a *causal* effect on candidate persistence, policymakers may want to address this issue by explicitly promoting more women as candidates. 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 intrinsically important (e.g., for downstream consequences of diminished female representation). In sum, we may have reason to care about both the descriptive heterogeneity in effects (what is captured by what we call the HiD) as well as the causally interpretable moderation in effects (what is captured by what we call the MiD). Crucially, the two quantities are important for different reasons, and so we need to be careful in distinguishing them.

Another application of RD effect conditionality studies how the timing and nature of elections affect the magnitude of incumbency advantages. We might infer from the difference in incumbency advantages across parties (Eggers, 2015), election cycles (de Benedictis-Kessner, 2018), economic or geographic contexts (Novaes and Schiumerini, 2021), 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: If reformers or policymakers 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 another mechanism – for example, certain parties or off-cycle elections attracting lower-quality candidates.

These examples illustrate the importance of carefully distinguishing between the HiD and MiD. 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 variation how carefully researchers distinguish between the descriptive difference in effects between subgroups, and a causal interpretation of said difference.

2.2 Related Work

This study relates to a corpus of work in which researchers have investigated RD effect conditionality under the banner of the term “difference-in-discontinuities.” However, the term “difference-in-discontinuities” has been used in various and inconsistent ways. The term has alternatively been used to describe either an estimand or an estimation approach, with studies often not being explicit about this distinction. Even more confusingly, the term has also been used to refer to entirely different underlying estimands of interest.

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, with the subsets defined by particular background characteristics (e.g. Desai and Frey, 2021; Lalive, 2007; Micozzi and Lucardi, 2021) or contexts (e.g. Becker et al., 2013; Card and Giuliano, 2016). This relates directly to research on treatment effect heterogeneity or conditional average treatment effects in RD designs (e.g. Becker et al., 2013; Hsu and Shen, 2019). In a second way in which the term has been used, 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), where those time periods are associated with some other intervention. Finally, a third use case refers to attempts to incorporate a more extensive panel data structure, along the lines of generalized difference-in-differences methods, into the RD design (Olson, 2020). Given the contested and inconsistent usage of “difference-in-discontinuities” in the literature, we opt to avoid this term throughout this paper. In Appendix B.2, we provide a lengthier discussion of the relationship between our quantities of interest and the different variants of “difference-in-discontinuities.”

Other related work touches on important issues concerning the causal role of variables other than the standard treatment in RD designs. Marshall (2019) discusses the use of close election RD designs to identify the effect of candidate-related attributes at the cutpoint other than the treatment, and questions its validity due to inferential and estimation challenges of isolating the causal effect of a variable at the cutpoint. Feigenbaum et al. (2017) introduce 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). However, their approach is quite specific to the context of studying majority-party control, rather than serving as a more general framework. Finally, Jenkins et al. (2016) combine propensity score estimation with a design focused on RD effect heterogeneity. While the econometric specification is explicated, the precise estimand of interest is less clear and never formalized; nonetheless, their strategy appears to be very much in the spirit of what we call the MiD.

3 NOTATION AND SETUP

As in a standard formalization of the RD design (e.g. Imbens and Lemieux, 2008), this study employs the potential outcomes framework of Neyman (1923) and Rubin (1974) to define causal effects. 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. 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 $Y_i(0, 0)$, $Y_i(1, 0)$, $Y_i(0, 1)$, and $Y_i(1, 1)$.

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 called the running (or forcing) variable and denoted here by $X_i \in \mathbb{R}$ such that $T_i = \mathbf{1}(X_i > c)$ for some cutoff $c \in \mathbb{R}$.

$S_i \in \{0, 1\}$ 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. In addition, $\boldsymbol{W}_i \in \boldsymbol{\mathcal{W}}$ will denote a vector of other observed pre-treatment covariates, which we will refer to as the "control set."[2]

Finally, precision in our 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 (i.e. moderator) 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). 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.

[2] The inclusion of this control set relates to other methodological research on the inclusion of covariates in RD designs (Calonico et al., 2019; Frölich and Huber, 2019). The use of covariates proposed and formalized in these studies contrasts with the present purposes, where these 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.

4 HETEROGENEITY-IN-DISCONTINUITIES

4.1 Estimand

Using the notation and setup above, the first estimand that we define is what we call the Heterogeneity-in-Discontinuities (HiD):

Definition 1 (Heterogeneity-in-Discontinuities (HiD))

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

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

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 $Y_i(t)$. This "reduced form" version of the HiD is thus $E[Y_i(1)-Y_i(0)|X = c, S_i = 1]-E[Y_i(1)-Y_i(0)|X = c, S_i = 0]$. Nonetheless, it is worthwhile considering the full definition as it highlights the role that S does (and does not play) and will allow for a more direct comparison with the MiD later. Specifically, given the purely observed role of S, the HiD represents an observed difference in RD effects across the two subsets $S = 1$ and $S = 0$, without establishing any causal influence on the part of S. This role of S also means that the effect of S on X need not be considered in the HiD (in contrast to the MiD, as will be discussed later).

The HiD falls directly under the umbrella of what has been discussed as treatment effect heterogeneity or conditional average treatment effects in RD designs (Hsu and Shen, 2019). For applied researchers conducting analysis of conditional RD effects by estimating separate RD effects across different subsets—perhaps the most common route for analyzing RD effect conditionality in applied work—their analysis maps directly onto the HiD, whether this is intended or not.

4.2 Identification and Estimation

Identification of the HiD 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 $s \in \{0,1\}$.

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 $s \in \{0,1\}$:

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

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

$$E[Y_i(0, S_i)|X_i = c, S_i = s] = \lim_{x \uparrow c} E[Y_i|X_i = x, S_i = s] \quad (2)$$

Hence, the HiD is identified as such:

$$\begin{aligned} & E[Y_i(1,S_i) - Y_i(0,S_i)|X=c, S_i=1] - E[Y_i(1,S_i) - Y_i(0,S_i)|X=c, S_i=0] \\ = & \left(\lim_{x \downarrow c} E[Y_i|X_i=x, S_i=1] - \lim_{x \uparrow c} E[Y_i|X_i=x, S_i=1]\right) \\ - & \left(\lim_{x \downarrow c} E[Y_i|X_i=x, S_i=0] - \lim_{x \uparrow c} E[Y_i|X_i=x, S_i=0]\right) \end{aligned} \tag{3}$$

This is tantamount to identifying two separate RD-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 HiD 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 HiD 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:

$$\begin{aligned} E[Y_i|S_i, X_i] &= \alpha_0 + \alpha_1 \tilde{X}_i + \alpha_2 S_i + \alpha_3 \tilde{X}_i \cdot S_i + \\ & \quad \beta_0 T_i + \beta_1 T_i \cdot \tilde{X}_i + \beta_2 T_i \cdot S_i + \beta_3 T_i \cdot \tilde{X}_i \cdot S_i \end{aligned} \tag{4}$$

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, β_2 identifies the HiD. 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 HiD's focus on differences in the RD treatment effect 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. In other words, rather than knowing whether a difference in RD effects simply manifests observationally across values of S, the question is whether S itself *causes* a difference in RD effects. Here, we define the Moderation-in-Discontinuities (MiD) to capture this phenomenon. As will be seen throughout this section, there are nontrivial complications involved in interpreting, identifying, and estimating the MiD, relative to the HiD.

5.1 Estimand

The MiD is defined as follows:

Definition 2 (Moderation-in-Discontinuities (MiD))

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

Several remarks are in order to highlight how the MiD differs from the HiD, understand how to correctly interpret the MiD, and clarify why the MiD is defined as it is.

To begin, notice that the MiD and HiD both have a similar structure that employs a difference-in-differences with respect to expectations of the potential outcomes $Y(t,s)$. However, the two quantities deviate fundamentally from one another by virtue of what the conditional expectation functions (CEFs) in each of them do (or do not) condition upon. In the case of the HiD, the CEFs condition upon $X(S) = X = c$, as in the standard RD framework. Recall that this denotes extrapolating the (smooth) CEFs to the cutpoint, rather than conditioning upon a (potentially non-existent) subpopulation for whom $X = c$. In addition, each CEF in the HiD also conditions upon S (a discrete moderator) and is hence defined with respect to a subpopulation within the larger population of interest. It is for these reasons that the HiD represents the difference in RD effects across two (observed) subpopulations.

In contrast, the CEFs in the MiD differ in two ways. First, they do not condition upon S, and hence are taken with respect to the entire population of interest. On this basis, the MiD (in contrast to the HiD) captures a causal influence of S; it represents how the RD effect (i.e. effect of the treatment on the outcome at the cutpoint) would change if *all* units had their moderator values S set to 1 vs. set to 0. For this reason, it must also be that the moderator of interest is itself theoretically mutable at the level of individual observations. This is in contrast to the HiD, and it should be one of the first considerations for applied researchers when deciding on their target estimand.[3]

Second, the CEFs in the MiD are different from the HiD in that they condition upon $X(s)$ rather than simply X. This is a critical detail in the MiD definition. As already noted above, the setup in this paper allows for the possibility of a causal effect of S on X, and hence one must consider the potential outcomes $X(s)$. The HiD only involves X because the estimand simply takes S as observed, and of course $X_i(S_i) = X_i$. In contrast, the focus on the causal influence of S in the MiD requires explicit conditioning upon $X(1)$ and $X(0)$ in the CEFs.

The definition of the MiD of course covers the special case where there is no causal relationship between S and X, in which case the estimand simplifies to $E[Y_i(1,1) - Y_i(0,1)|X = c] - E[Y_i(1,0) - Y_i(0,0)|X = c]$. This would follow 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), or in situations where

[3] 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's imagination to consider a counterfactual election with a candidate that was similar in all ways except of a different race.

there are conceptual or substantive reasons to believe that S does not affect X even if not randomized. When there is a causal relationship between S and X, however, not only must there be conditioning upon $X(1)$ and $X(0)$, but each CEF must condition upon the appropriate quantity of the two, and additional care must be taken in interpreting what the MiD represents. In particular, each of the four CEFs that the MiD can be disaggregated into can be represented as $E[Y_i(t, s)|X_i(s) = c]$, where (as can be seen) both potential outcomes $Y(t, s)$ and $X(s)$ within any single CEF are defined with respect to the *same* value s. That is, for each CEF, the expectation of the potential outcome $Y(t, s)$ for state s is taken conditional upon the potential outcome $X(s)$ defined in the same state s. Further, note that the first half of the MiD pertains to state s (i.e. 1) and the second half pertains to state s' (i.e. 0).

It may at first glance seem strange or unconventional that the MiD features different components that condition upon different potential outcomes. But for several reasons explained below, this actually makes the MiD meaningful from a conceptual perspective, as well as internally coherent from a theoretical perspective.

First, notice that this form for the MiD is consistent with the description and interpretation of the MiD presented above: the MiD represents the causal change in the RD effect that would be induced by switching units between states s and s'. If units were intervened upon and switched into a state s, this would have implications for both sets of potential outcomes—not only $Y(t, s)$ but also $X(s)$. In other words, since each component of the MiD (each CEF) pertains to a particular state s or s', that state must be held constant within the CEF. In such a way, the MiD as defined maps exactly onto the idea of counterfactual values for S while taking into account the fact that S can affect both Y and X. Indeed, the definition of the MiD is precisely what would manifest if one were to experimentally manipulate S and then compare RD effects across the two experimental groups.

Second, one might also consider the theoretical relevance of possible alternative definitions for a MiD-like estimand, which could seek to hold the quantities being conditioned upon fixed across CEFs. For instance, one might propose an alternative estimand that is the same as the MiD except that it conditions upon $X(0) = c$ for all CEFs. Not only would this not track with the intuition of the MiD as previously discussed, but it would feature a somewhat nebulous component that takes the expectation of a potential outcome $Y(t, s)$ that exists in one state of the world (s) conditional upon another potential outcome $X(s')$ that exists under an entirely different state of the world (s'). Such an endeavor may make sense in other contexts, such as with principal strata or in other settings where one can define units or subsets of units by specific/discrete potential outcome values. But recall that in this RD context, the conditioning upon the running variable does not represent conditioning upon a subpopulation but rather extrapolation to a point (and a point at which there may actually not be any units that exist). Theoretical coherence of the quantity in this context requires conditioning the expectation of $Y(t, s)$ upon $X(s)$ in the same state of the world; there is limited internal coherence to extrapolating a quantity under one state of the world with respect to another quantity that does not exist concurrently.

Conditioning only upon X is also not an option (unless S is known to not affect X) for this same reason. Consider that $E[Y_i(t, s)|X_i] = E[Y_i(t, s)|X_i(S_i)] = p(S_i = s)E[Y_i(t, s)|X_i(s), S_i = s] + p(S_i = s')E[Y_i(t, s)|X_i(s'), S_i = s']$. Hence, in situations where S affects X, the quantity $E[Y_i(t, s)|X_i]$ implicitly exhibits the same problem as described in the previous paragraph. In addition, as will be noted later, identification of the MiD will require a conditional independence assumption for the moderator. In situations where S affects X, X is by definition posterior to S, and hence conditioning upon X is *counter* to the goal of establishing the conditional independence of the moderator (akin to conditioning on a post-treatment variable). In contrast, $X(s)$ and $X(s')$ are by definition prior to S.[4]

In sum, the MiD is defined in a way that coheres with the unique features of the RD

[4] Further, to the extent that a unit would select into a particular value of S, 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 $X(s')$.)

setting. That being said, these unique features at the same time result in a conceptual wrinkle: specifically, in cases where S affects X, because different counterfactual values of the moderator also imply different values of X, the units to the left (right) of the cutpoint (and similarly, the units that have a value of X within any specified distance from c) will not necessarily be the same units under the different counterfactual values of S. This further highlights the importance, as already noted, of interpreting the MiD as a moderation to the effect at the cutpoint, not the moderation of an effect overall on average, for any subset, or for any particular unit.[5] This is of course a natural extension of the localness of a standard RD effect.[6]

5.2 Identification and Estimation

The MiD requires more demanding identification assumptions and accompanying estimation effort than the HiD. In this section, we focus on one particular strategy for identifying and estimating the MiD that is most pertinent in typical RD settings, in which there has not been randomization of the moderator. In the Appendix Section A.2, we also present alternative identification strategies that may be applicable in some cases under different sets of assumptions.

Like with the HiD, the identification strategy for the MiD 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)

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

[5] This also relates to 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 MiD presented here hence implicitly echoes some of Marshall's concerns.

[6] 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.

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 HiD represented in Assumption 1. As described earlier, the HiD's 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 MiD 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 this version of continuity is actually more demanding as a practical matter, as discussed in more detail in Appendix C.2.

In addition, identification of the MiD relies upon *conditional* independence of the moderator, along with an accompanying requirement for common support. Specifically, the following assumptions are made:

Assumption 3 (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 4 (Moderator Common Support) [7]

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

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

$$\begin{aligned} E[Y_i(1,s)|X_i(s)=c] &= E_W\left[E[Y_i(1,s)|X_i(s)=c,\boldsymbol{W}_i]\right] = E_W\left[E[Y_i(1,s)|X_i(s)=c,\boldsymbol{W}_i,S_i=s]\right] \\ &= E_W\left[E[Y_i(1,s)|X_i=c,\boldsymbol{W}_i,S_i=s]\right] = \lim_{x\downarrow c} E_W\left[E[Y_i(1,s)|X_i=x,\boldsymbol{W}_i,S_i=s]\right] \\ &= \lim_{x\downarrow c} E_W\left[E[Y_i|X_i=x,\boldsymbol{W}_i,S_i=s]\right] \qquad (5) \end{aligned}$$

[7] 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.

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

$$E[Y_i(0,s)|X_i(s)=c] = \lim_{x \uparrow c} E_W\left[E[Y_i|S_i=s, X_i=x, \boldsymbol{W}_i]\right] \tag{6}$$

The steps and conditions required for identification of the MiD are clearly more demanding than that for the HiD. If the moderator has not been randomized, the assumption of conditional independence of the moderator is the core assumption upon which identification of the MiD 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, 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 MiD 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 other words, even if one is uncomfortable fully embracing the conditional independence assumption, one can still view an evaluation of the MiD under this framework 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 reasoning 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.

Under these appropriate cautions, researchers can implement the covariate adjustment that is necessary for estimation by building upon the local linear regression framework that is commonly used to estimate simple RD effects. Below (and in the Appendix), 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:

$$\begin{aligned} E[Y_i|S_i, X_i, \boldsymbol{W}_i] \;&=\; \alpha_0 + \alpha_1\tilde{X}_i + \alpha_2 S_i + \alpha_3\boldsymbol{W}_i + \alpha_4\tilde{X}_i \cdot S_i + \alpha_5\tilde{X}_i \cdot \boldsymbol{W}_i + \\ &\quad\; \beta_0 T_i + \beta_1 T_i \cdot \tilde{X}_i + \beta_2 T_i \cdot S_i + \beta_3 T_i \cdot \boldsymbol{W}_i + \beta_4 T_i \cdot \tilde{X}_i \cdot S_i + \beta_5 T_i \cdot \tilde{X}_i \cdot \boldsymbol{W}_i \end{aligned} \tag{7}$$

where $\tilde{X}_i = X_i - c$. Under this model, β_2 identifies the MiD.

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 MiD and its identification. In contrast, prior research attempting to recover estimates of a MiD 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). In Appendix Section A.1, we also discuss how additional higher-order interactions might be included as well as regularization penalties incorporated into the estimation.

Local Linear Regression with a Matched Sample. Choices such as that between model (7) above and a more highly interactive model (such as model (A.1) in the Appendix) 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 could help 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 samples.

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, one can estimate the MiD with local linear regression using equation (4), i.e. the same model used to estimate the HiD. Applied to a matched sample, however, β_2 recovers the MiD.

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 allows 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 the MiD for the moderated (i.e. for units for which $S = 1$), or MiD for the unmoderated (i.e. for units for which $S = 0$).

6 GUIDANCE FOR APPLIED RESEARCHERS

Next, we offer a set of "best practices" for applied researchers whose empirical work centers on conditional RD effects. Regardless of the estimand of interest, we suggest that researchers:

1. carefully distinguish between HiD and MiD as distinct estimands; specify and justify which of those is the project's quantity of interest.

2. check estimates' sensitivity to different bandwidth choices. This is particularly important in the case of the MiD, for which no formalized optimal bandwidth estimator exists. In the case of the HiD, standard procedures could be employed to determine the MSE-optimal bandwidth separately across the two subsets of S, or the larger of the two could be employed if a single estimating model is used. In the case of the MiD, we recommend (as a principled though not formalized approach) employing the same bandwidth selection as would be used for the HiD and subsequently checking robustness to a range of other bandwidths.

If the target estimand is the HiD, researchers should proceed with estimation as usual but remain careful and explicit about the non-causal role of the moderator. If the target estimand is the MiD, we advise that researchers additionally:

3. discuss the plausibility of the conditional independence assumption in the context of the research design and available control set. Even where it cannot be fully satisfied, applying our framework with a relevant control set offers a principled way of distinguishing between alternative, ex-ante theoretically justified mechanisms, and strengthening confidence in the role of the theorized moderator in accounting for the difference in RD effects.

4. evaluate common support between moderated and unmoderated units. If there is no (or poor) overlap, the MiD cannot be recovered (Appendix C.1).

5. demonstrate continuity of variables in the control set around the threshold following existing RD diagnostics (Appendix C.2).

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. Table A.1 in the Appendix summarizes these works. Overall, our review suggests the following:

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

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

3. Papers rarely report any kind of overlap analysis.

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

We illustrate our framework in greater detail using four papers where we have been able to replicate the research design and extend it with a sufficiently rich set of controls, $\boldsymbol{W}$. Of these, we discuss Cipullo (2021) and Desai and Frey (2021) in-depth below and discuss the other two in Appendix E, highlighting additional issues concerning the evaluation of plausibly exogenous moderators and RD effect heterogeneity with respect to time.

7.1 Gender Gap in Winning On Persistence

Are women less persistent in running for office when losing an election? A line of recent work (Bernhard and de Benedictis-Kessner, 2021; Wasserman, 2018, 2021) studies this question in the context of the United States employing a HiD design. Outside of the U.S., Cipullo (2021) uses the same design to study the gender gap in persistence in Italian mayoral candidates. We replicate and extend Cipullo (2021), reporting estimates of both the HiD and MiD.[8]

7.1.1 Original Design and Interpretation

Cipullo (2021), following Wasserman (2018) and Bernhard and de Benedictis-Kessner (2021), studies whether the attrition effect of losing an election on the probability to run again is differentially greater for female candidates in U.S. House and Italian mayoral races. There are multiple reasons why women might suffer from greater attrition – only some of which are directly attributable to gender. As discussed earlier in the paper, this distinction between possible mechanisms is an exemplary case for the differentiation between the HiD and MiD. It underlines the importance of carefully specifying the theorized role of gender, which in turn has implications for which estimand is the relevant quantity of interest.

Though not explicitly spelled out, we consider the implied estimand of the paper to be the MiD. Cipullo (2021) applies the design in order to examine the "sticky floor hypothesis", which states that underrepresented groups 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 the paper's theoretical interest in the MiD, Cipullo (2021)'s key empirical specification does not exactly match either the HiD or the MiD estimator (although it may come close to the HiD). The specification uses candidates' margin of victory as the running variable, and assigns the treatment to candidates who barely lost; in addition to the interaction between treatment and running variable, the treatment and running variables are also interacted with a dummy for female candidates (as well as including the triple interaction) in order to capture the heterogeneity. [9] As a departure from the HiD estimating equation (4),

[8] We collected data on Italian municipal elections and merged it with the Italian Ministry of Interior's Dataset on Elected Officeholders. Small discrepancies with Cipullo (2021) may remain as we did not have access to the author's original replication data.

[9] Cipullo's formal specification includes $f(F_i; WinningMargin_{i,t})$. but does not explicitly state interactions

Cipullo (2021) also uses election year fixed effects (without any interaction). The addition of year fixed effects means that the proposed specification does not map cleanly onto either the HiD or the MiD. Overall, the discrepancy between the estimand of interest (MiD) and the estimator is illustrative of many surveyed papers.

7.1.2 Applying Our Framework

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

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 D.1 reports propensity scores before matching). We observe good overlap across bandwidth choices. This indicates that male and female candidates in our data have broadly similar characteristics, which is desirable, and allows us to proceed with estimation of the MiD.

Robustness to MiD Estimators. Next, we assess whether the gender gap changes in magnitude when we estimate the MiD. Figure 2 reports the estimates at different bandwidths using the HiD estimating equation (4) along with the various MiD estimators discussed earlier. We see that, overall, the magnitude of the estimate does not change significantly

between the treatment and the running variable. Given references to Wasserman (2018), we believe that Cipullo's notation intends to subsume the interactions with the treatment in this term.

across specifications: the gender difference in the effect of winning on running again persists when accounting for the control set.

We emphasize that, given the limited control set, we should be careful in interpreting the gender gap as a causal effect of gender. Nonetheless, applying our framework strengthens confidence in the results and can help to credibly rule out some mechanisms that rely on gender correlates (e.g., age). 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 MiD.

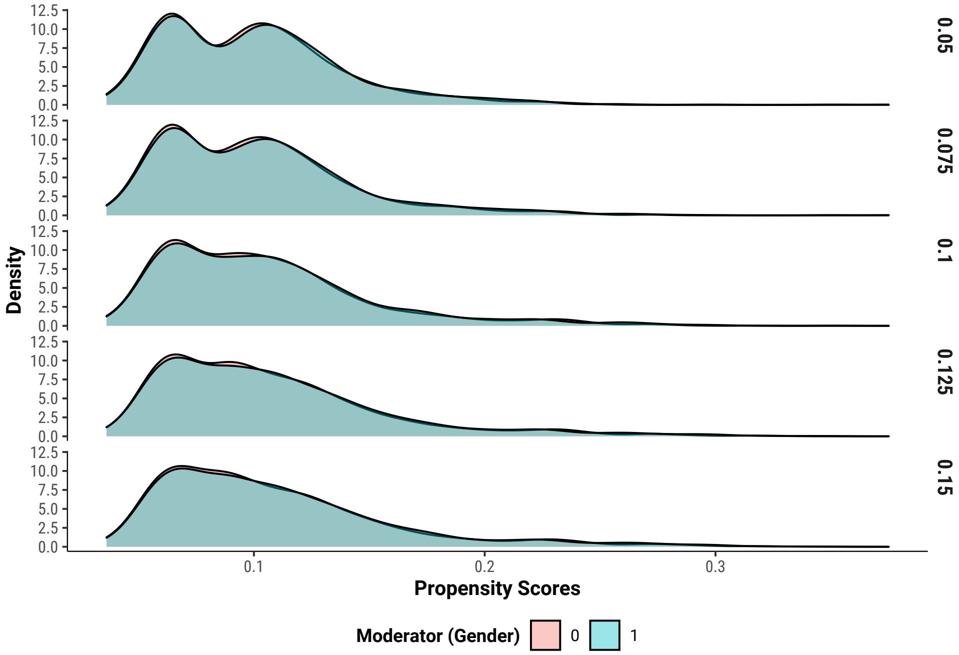

Figure 1: Distribution of moderator propensity scores in Cipullo (2021) after matching.

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). We discuss the ambiguity in the intended estimand, and estimate a meaningful difference between the

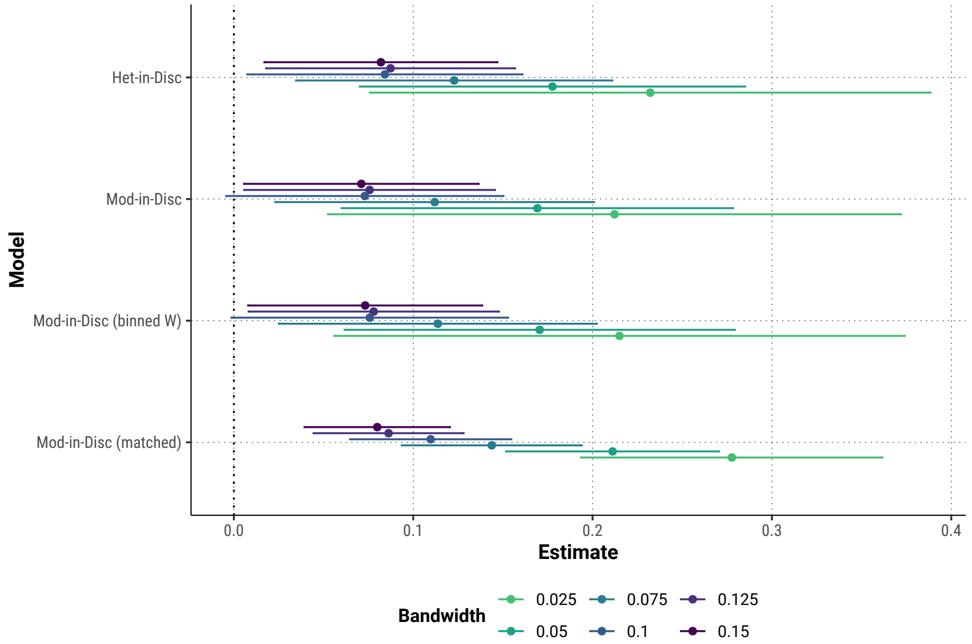

Figure 2: MiD estimates of the moderation effect of candidate gender on incumbency advantages, following Cipullo (2021)

HiD and the MiD. 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.

7.2.1 Original Design and Interpretation

Desai and Frey (2021) are interested in 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. The authors hypothesize 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 also differ with respect to other characteristics, such as geography, climate, or demographics. If that is the case, we may observe a policy equivalence in high-poverty areas due to characteristics merely associated with poverty; for example, because a more remote location requires greater spending on sanitation basics irrespective of being targeted towards poverty.

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 MiD – the moderating effect of poverty 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 across the two subsamples (high- and low-poverty towns), which maps onto the HiD.

The authors then use a specification that is similar to, but does not fully map onto the HiD. 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 contest- and town-level covariates as linear regressors (without interactions). These additional parameters are not formally justified and leave the implemented estimator without a mapping to a clearly defined estimand. We also highlight that the authors report results for a limited selection of bandwidths (see Appendix D.2).[10]

7.2.2 Applying Our Framework

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

Control Set. We focus on the following municipality-level controls that are susceptible to feature an independent moderation effect on winners' policy decisions: past budget size, the vote share the left obtained in past presidential elections, inequality, GDP per capita,

[10] They also use triangular kernels throughout; in Appendix D.2 we show that the original specification (at the reported bandwidths) yields similar results when using uniform kernels instead.

population (logged), longitude, and latitude.[11] We use this control set to assess overlap and estimate MiD effects.

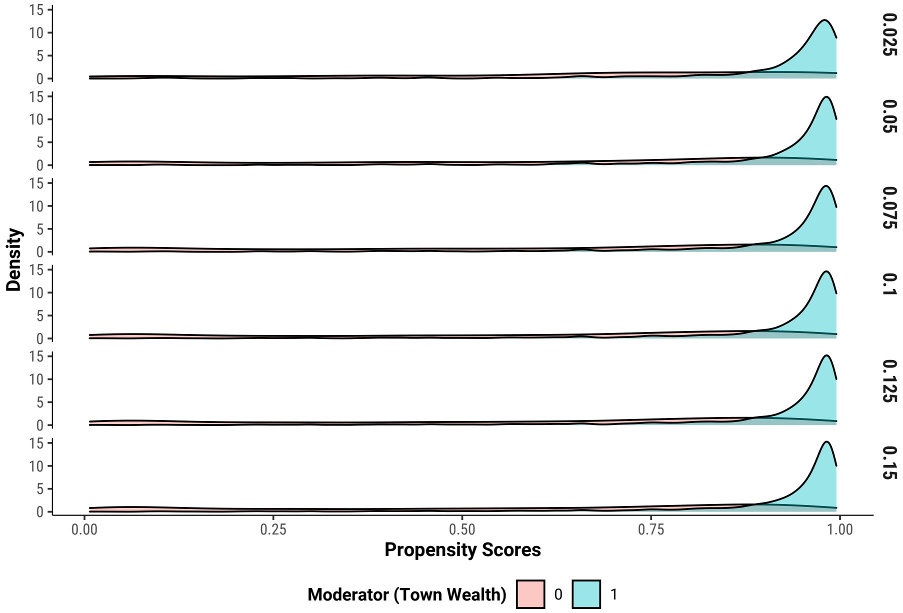

Figure 3: Distribution of moderator propensity scores in Desai and Frey (2021) after matching.

Overlap Assessment. Figure 3 reports the distribution of propensity scores for high-poverty (unmoderated) and low-poverty (moderated) municipalities after matching with replacement (pre-matching propensity scores are in Appendix D.1). Even when 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. This application highlights the importance of the overlap check to assess the plausibility of alternative moderating variables that may account for the contrast.

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

[11] Appendix D.3 shows similar results with only longitude, latitude, and population in the control set.

if the overlap is as sparse as in this case. Figure 4 presents our estimates. Our HiD estimate matches the original paper's results close to the authors' preferred bandwidth (0.052), though our confidence intervals are larger due to the omission of linearly added control variables and year fixed effects. As we move towards wider bandwidths, however, the HiD estimates attenuate towards zero.

In local linear regression estimates of the MiD using equation (7), 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 estimating the MiD using matching (without replacement), we get results that are substantively similar to the original HiD; 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.

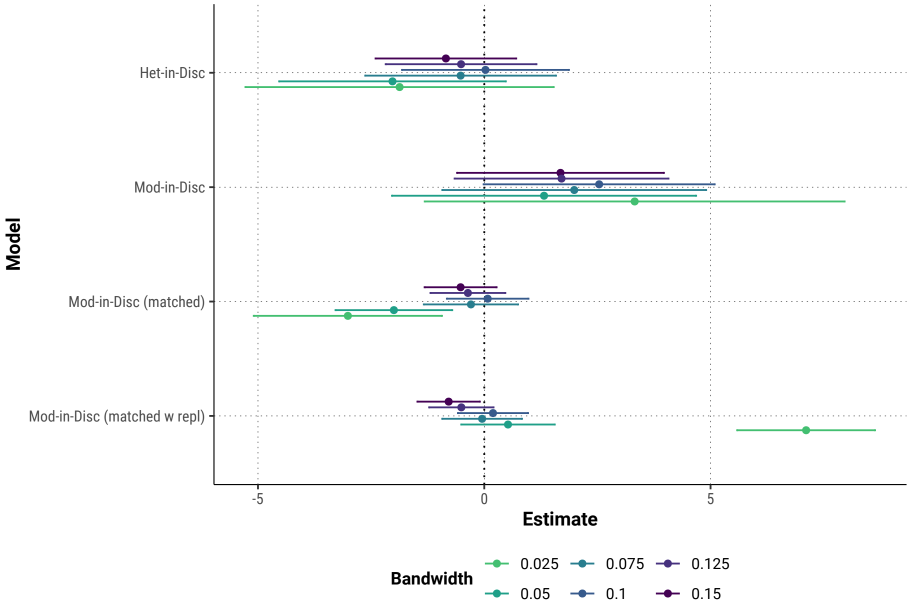

Figure 4: MiD estimates of the moderation effect of low town poverty on the effect of a right-wing mayor on pro-poor policies, following Desai and Frey (2021).

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 overlap with respect to the moderator in order to justify any credible interpretation of estimates as pertaining to the MiD. Third, even aside from overlap concerns, the MiD estimates may be fundamentally different from the HiD ones, casting doubt on causal interpretations of the moderation 'effect' when using the HiD specification. In this particular application, the results suggest we cannot rule out the possibility that a factor correlated with town wealth (e.g., geography) is responsible for the observed heterogeneity.

8 CONCLUSIONS

In this paper, we introduce and discuss an important distinction between two quantities of interest when assessing effect conditionality in regression discontinuity designs: (1) the HiD, which maps onto differences in the RD effect conditional on a given characteristic, and (2) the MiD, which recovers the moderation effect of the defining characteristic that causes the RD effect to change. Separating the two estimands is not only important for conceptual reasons, but can also help researchers and policymakers draw more precise conclusions from their findings.

We present a formalized framework to describe and differentiate between the two estimands in detail, and we offer strategies to identify and estimate each of them. 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 work comes with an important caveat. We stress that in many settings, the conditional independence assumption necessary for identifying the MiD may not be credible, or gathering an appropriate control set may not be feasible. Despite these challenges, our framework introduces new methodological vocabulary, important conceptual distinctions, and useful estimation strategies that can benefit researchers across many social science settings where RD effect conditionality is of interest.

References

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

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).

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.

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*.

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

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

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.

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

Grembi, V., Nannicini, T., and Troiano, U. (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics*, 8(3):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.

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.

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

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.*

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.

Novaes, L. M. and Schiumerini, L. (2021). Commodity shocks and incumbency effects. *British Journal of Political Science*, pages 1–20.

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.

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.

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. *The Review of Economics and Statistics*, pages 1–46.

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

Appendix for

Effect Heterogeneity and Causal Attribution in Regression Discontinuity Designs

Table of Contents

A ADDITIONAL METHODS

A.1 MiD Local Linear Regression with Regularization

Given that the baseline specification presented in the main text for estimating the MiD (given a conditionally independent moderator) already contains interactions between continuous X and potentially continuous variables contained in $\boldsymbol{W}$, one might be concerned about noise and sensitivity due to over-parameterization. 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 (ℓ^1, ℓ^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:

$$\begin{aligned} E[Y_i|S_i, X_i, \boldsymbol{W}_i] \quad = \quad & \alpha_0 + \alpha_1 \tilde{X}_i + \alpha_2 S_i + \alpha_3 \boldsymbol{W}_i + \alpha_4 \tilde{X}_i \cdot S_i + \alpha_5 \tilde{X}_i \cdot \boldsymbol{W}_i + \\ & \alpha_6 S_i \cdot \boldsymbol{W}_i + \alpha_7 \tilde{X}_i \cdot S_i \cdot \boldsymbol{W}_i + \\ & \beta_0 T_i + \beta_1 T_i \cdot \tilde{X}_i + \beta_2 T_i \cdot S_i + \beta_3 T_i \cdot \boldsymbol{W}_i + \beta_4 T_i \cdot \tilde{X}_i \cdot S_i + \beta_5 T_i \cdot \tilde{X}_i \cdot \boldsymbol{W}_i + \\ & \beta_6 T_i \cdot S_i \cdot \boldsymbol{W}_i + \beta_7 T_i \cdot \tilde{X}_i \cdot S_i \cdot \boldsymbol{W}_i \end{aligned} \tag{A.1}$$

In this case, the MiD is identified by $\beta_2 + \beta_6 E[\boldsymbol{W}_i]$.

A.2 MiD Alternative Identification Strategies

Identification of the MiD could also proceed from alternative strategies with alternate sets of assumptions, which may be applicable in certain research contexts. Each of these alternatives proceeds from a different statistical design/model setup and implies different estimation procedures.

First, the most theoretically straightforward strategy to identify and estimate the MiD relies upon a moderator that has been randomized (or can otherwise be assumed to be unconditionally exogenous). Appendix Section A.2.1 formalizes and provides details on this strategy, showing how the MiD is equivalent to the HiD in this special case.

Second, another alternative involves deviating from the standard RD formulation based upon continuous conditional expectation functions. This strategy instead relies upon a design-based perspective 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). Appendix Section A.2.2 describes this alternative "local randomization" strategy in more detail. The benefit of this 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.

A.2.1 Exogenous Moderator

In addition to assumption 2, 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 5 (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 MiD is identified as such:

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

Note that the quantity that identifies the MiD in equation (A.2) is precisely the same quantity that identifies the HiD in equation (3). Hence, the implication is that, given the special case of randomization of the moderator, the HiD is indeed the MiD and can be interpreted as such.[12]

Given that the quantity that identifies the HiD also identifies the MiD when the moderator has been randomized, the estimation procedure can also follow that used in the HiD 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, β_2 now identifies the MiD of interest. As before, estimation can be done via a local linear regression.

A.2.2 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 MiD would proceed with the following local randomization assumption, adapted from Mattei and Mealli (2016):

Assumption 6 (Local Randomization of Running Variable) [13]

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 $s \in \{0, 1\}$.

[12] As will be described in a later section, 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. Such an approach is tantamount to assuming the policy (S) is randomly assigned, in which case the HiD is the MiD; however, the plausibility of such an assumption in the case of policy change over time is a separate question.

[13] 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

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 $i \in \mathcal{U}_c$.

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 MiD need not be estimated on the basis of parameterizing 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 MiD estimate (see Bansak, 2021, for more details).

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).

B EXISTING LITERATURE

B.1 Literature Review

Table A1 summarizes recent studies using Heterogeneity-in-Discontinuities (HiD) designs.

Authors	Journal	Treatment	Outcome	Heterogeneity Set
Lindo et al. (2010)	AEJ:Applied	Academic Probation	Graduation	Demographics
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
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)	QJPS	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
Novaes and Schiumerini (2021)	BJPS	Being Elected	Winning Again	Commodity Shocks
Brown et al. (2020)	WP	Being Elected (State)	Being Elected (Congress)	Gender
Cipullo (2021)	WP	Being Elected	Running Again	Gender
McCrain and O'Connell (2022)	WP	Being Elected (State)	Being Elected (Congress)	Professionalisation

Table A1: Summary of recent papers using heterogeneity-in-discontinuity designs

B.2 Relationship with "Difference-in-Discontinuities"

As described in the main text, the term "difference-in-discontinuities" has been used in contested and inconsistent ways by researchers investigating RD effect conditionality. Here, we provide a lengthier discussion of the term, characterize (and distinguish between) the different ways in which the term has been used, and describe how these variants 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; Micozzi and Lucardi, 2021). 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 HiD presented and formalized in the present study.

A related concept by Becker et al. (2013) that maps onto the just mentioned use of 'difference-in-discontinuities', termed 'heterogeneous local average treatment effect' (HLATE), picks up the heterogeneity in RD effects at the cutpoint across the domain of continuous moderating covariates. The paper spells out an estimand that resembles our HiD, yet there are a number of differences and areas of ambiguity. First, the setup in Becker et al. (2013) does not define potential outcomes with respect to the conditioning variable and, as a related matter, remains ambiguous about how exactly to interpret the proposed estimand. This can have unwelcome consequences, such as applied researchers interpreting the proposed estimand too strongly as a causal quantity (see, for example Casarico et al. (2022)). Second, their identification strategy imposes an exogeneity assumption for the moderator that, as our framework shows, is not necessary for a descriptive Heterogeneity-in-Discontinuities. If one wishes to identify the Moderation-in-Discontinuities instead, then Becker et al. (2013) do not offer a way forward if the moderator is only *conditionally* independent, which is likely to be a more common setting than strict exogeneity. Finally, their proposed regression estimator does not include an interaction between the running variable and the moderator, which implies that the slopes of the CEF away from the cutoff are independent of the moderator. Our framework helps to address these issues more fully by explicitly distinguishing between and interpreting the HiD and the MiD as separate estimands defined by an expanded potential outcomes notation, as well as offering formally motivated estimators for each.

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 (Köppl-Turyna and Kantorowicz, 2020; Lassébie, 2020). 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 such a way, this variant of difference-in-discontinuities may represent a somewhat ambiguous, informal amalgam of the HiD and MiD. 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 (Olson, 2020). 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, in the spirit of the MiD. 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.

As made clear in the formalization presented in the main text, both the HiD and MiD have been 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, the present study 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. A rigorous formalization along those lines would be valuable future work.

C CAUTIONS FOR APPLIED RESARCHERS

C.1 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 MiD, and checking for common support should be a virtually automatic first step for researchers considering an investigation of the MiD. 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 MiD 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 MiD, it is either not identified or otherwise not plausibly estimable with the available data.

C.2 Continuity

As noted earlier, the assumption of continuity required for the HiD 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 MiD 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 HiD and MiD 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 HiD or MiD, 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, HiD, and MiD) 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).

D ADDITIONAL ROBUSTNESS CHECKS

D.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.

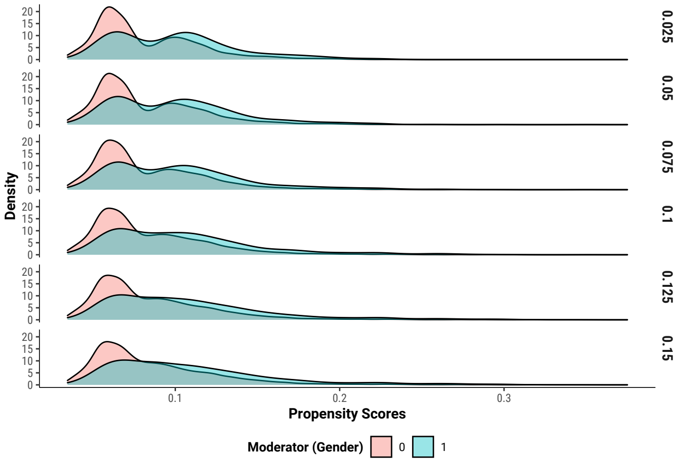

Figure A1: Assessing overlap among observations before matching in Cipullo (2021) using propensity scores (measuring how likely an observation is in the moderated group)

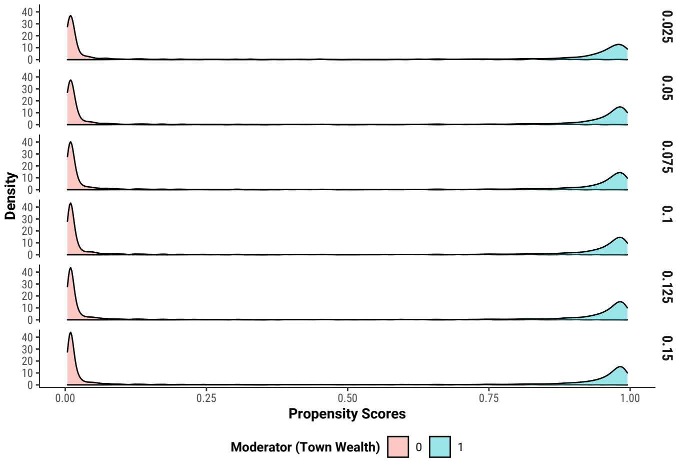

Figure A2: Assessing overlap among observations before matching in Desai and Frey (2021) using propensity scores (measuring how likely an observation is in the moderated group)

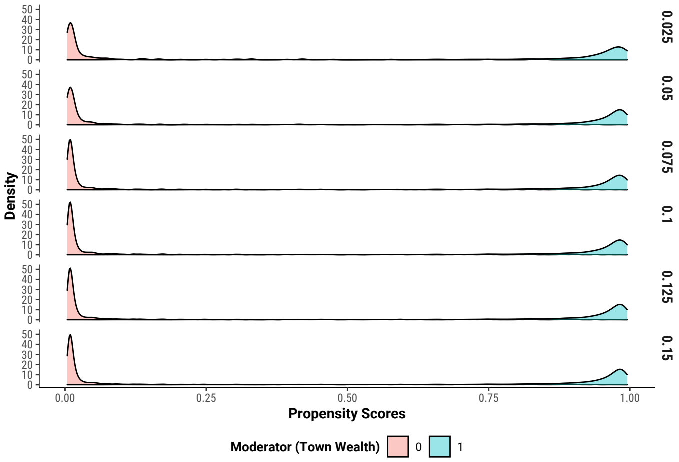

Figure A3: Assessing overlap among *matched* observations (without replacement) in Desai and Frey (2021) using propensity scores (measuring how likely an observation is in the moderated group)

D.2 Desai and Frey (2021): Sensitivity to Kernel and Bandwidth Choice

Throughout our applications, we use uniform kernels around the threshold in estimating the HiD and MiD. In the original paper, Desai and Frey (2021) use triangular kernels as the default estimation procedure. In Figure A4, we report our estimates from a specification tracking the original paper (with covariates and year fixed effects) for both kernel estimators across a wide range of possible bandwidths.

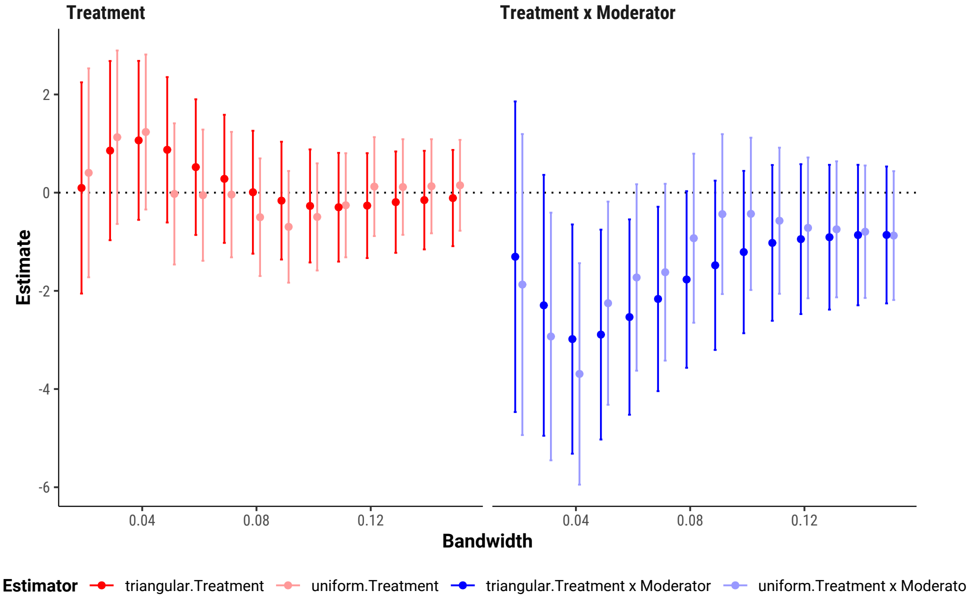

Figure A4: Assessing overlap among matched observations (without replacement) in Desai and Frey (2021) using propensity scores (measuring how likely an observation is in the moderated group)

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

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

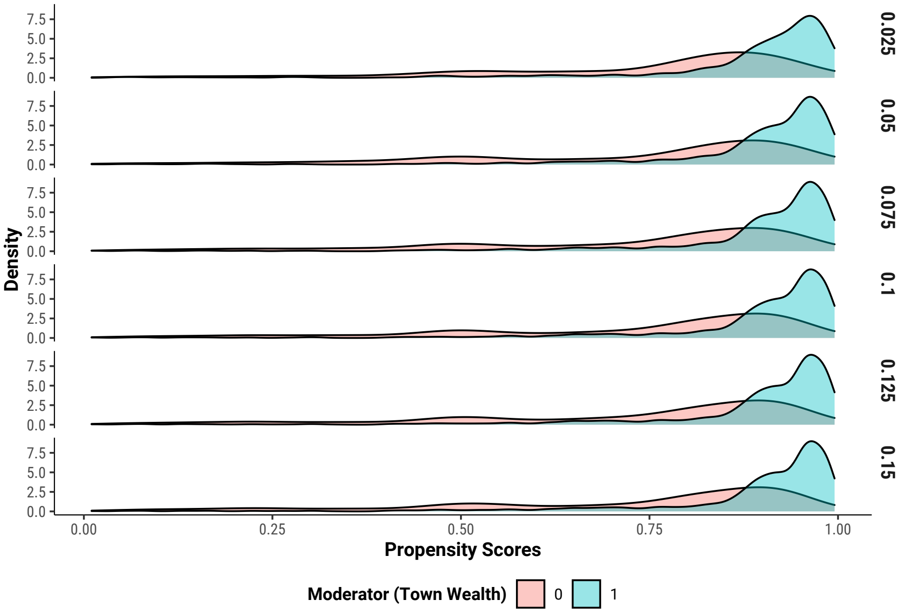

Figure A5: Assessing overlap among matched observations (with replacement) in Desai and Frey (2021) using propensity scores (measuring how likely an observation is in the moderated group) and a limited control set

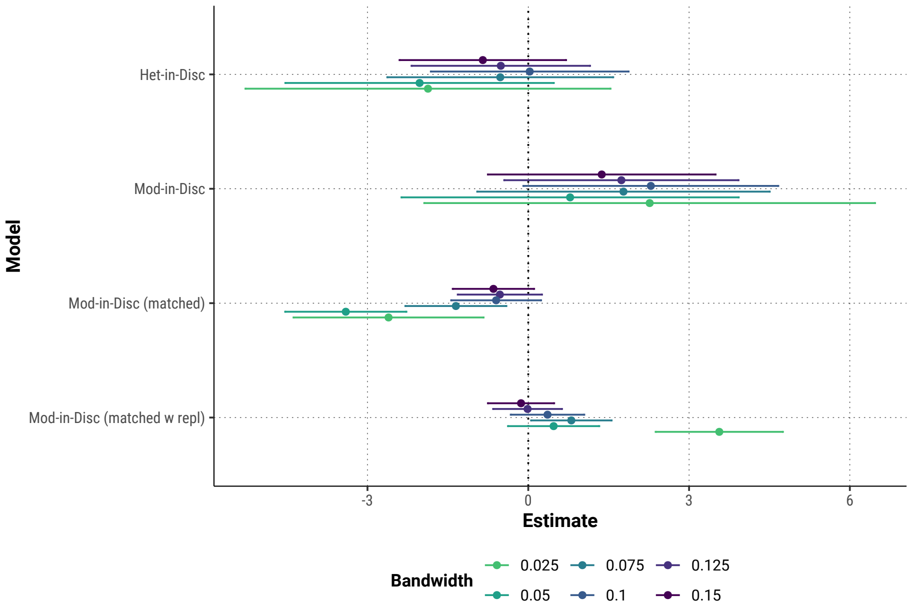

Figure A6: Moderation-in-Discontinuities estimates on the moderation effect of town poverty on the effect of electing a right-wing mayor on pro-poor policies, following Desai and Frey (2021) and using a limited control set.

E ADDITIONAL APPLICATIONS

In this section of the Appendix, we offer two additional applications that illustrate common issues with studies estimating conditional effects in an RD setup.

E.1 The Effect of Commodity Shocks on Incumbency Advantages (Novaes and Schiumerini, 2021)

This application examines the conditionality of incumbency (dis)advantages. In the context of Brazil, researchers have consistently documented that winning an election *reduces* the (now incumbent) candidate's probability of winning in the next cycle. This effect is far more pronounced in rural areas. Novaes and Schiumerini (2021) argue that this effect conditionality is not *because of* rural characteristics but, instead, caused by price shocks (inflation) in towns with these characteristics. Although this paper makes the strongest claims about the causal nature of the moderation effect amongst the set of our applications, the application of our framework raises questions about the moderator being "plausibly exogenous", especially as HiD and MiD estimates differ markedly. This underscores the importance of distinguishing between the two estimands both theoretically and empirically.

E.1.1 Original Design and Interpretation

Novaes and Schiumerini (2021) motivate their project by arguing out that while Brazilian incumbency disadvantages are significantly more severe in rural areas, this effect conditionality is likely spurious: instead, "exogenous price volatility has a strong impact on the electoral returns to incumbency" (p. 9). In other words, Novaes and Schiumerini (2021) suggest that while there are several municipal-level characteristics that are perhaps correlated with each other, the key variable with a *causal* moderation effect on incumbency advantages is price shocks. The authors make their interest in the MiD explicit at the beginning of the results section, where they "analyze the effect of price shocks on incumbents' electoral fortunes vis-a-vis challengers" (p. 9). This MiD is an important quantity in this context: if we believe that incumbency disadvantages may be harmful to accountability and incentivize officeholders to slack, then we need to understand the root causes behind this effect.

How can we distinguish between the moderation effect of price shocks versus other attributes of municipalities with harsher incumbency disadvantages? Importantly, Novaes and Schiumerini (2021) assume that their moderator – price shocks – is "plausibly exogenous".[14] Under this assumption, the HiD and MiD estimators would become equivalent to one another as the conditional independence assumption is satisfied without covariate adjustment. Indeed, the authors then proceed with a specification close to what we call the HiD estimator, although they depart from our operationalization by adding year fixed-effects (without any interactions). As usual in incumbency effect RDs, the running variable represents candidates' voteshare, while the treatment is assigned to those candidates who win the election. Both running variable and treatment are interacted with a continuous measure of price shocks; the triple interaction is also included.[15]

To support their claim that the moderator is exogenously assigned, Novaes and Schiumerini (2021) run a placebo check wherein the incumbency effect for the election in $t + 1$ is moderated by a measure of a *future* price shock (in $t+2$). Their results suggest that future price shocks do not affect the magnitude of the present incumbency effect, which implies that time-invariant municipal characteristics are unlikely to be correlated with price shocks. However, this evidence cannot rule out the presence of time-variant correlates, such as GDP or population. If, as Table A2 suggests, price shocks are correlated with economic output within municipalities, then conditional independence may not hold as long as we cannot identify the precise causal mechanism; consequently, we are unable to interpret the RD interaction coefficient as a straightforward MiD estimate.

[14] "The second identification assumption is that the municipal commodity price index is exogenous to structural municipal characteristics." (p. 9)

[15] The decision to operationalize the moderator as a continuous variable in the original paper imposes additional structural assumptions on the nature of the RD conditionality (namely, that the moderation effect is linear). To avoid this assumption, and in order to retain the binary nature of the moderator in line with our framework, we dichotomize the moderator variable into whether the price shock was negative ($S = 1$) or positive ($S = 0$).

Table A2: Correlates of price shocks, within-municipality estimates

	Model 1	Model 2	Model 3	Model 4
Population (log)	-0.473	-0.029	-1.279	0.059
	(2.078)	(0.047)	(2.663)	(0.064)
Gini	-6.673	0.213	-3.974	0.204
	(3.409)	(0.129)	(4.040)	(0.165)
GDP (log)	6.117	-0.135	9.097	-0.174
	(0.879)	(0.021)	(1.058)	(0.029)
Outcome	Continuous	Binary	Continuous	Binary
N	42,439	42,439	22,033	22,033
Sample	Full	Full	Rural	Rural

All models include municipality and year fixed effects. Standard errors clustered by municipality in parentheses. 'Continuous' denotes a continuous inflation measure, while 'Binary' uses a binary outcome that is 1 if the inflation measure is negative. Columns with full sample report estimates fitted on all municipalities while 'Rural' reports estimates fitted on rural municipalities only.

E.1.2 Applying Our Framework

Control Set. We use a wide set of municipality-level attributes – log(population), GDP, inequality, GDP, share non-white population, share rural population, and the number of canddidates in the race. We use this control set to assess overlap and estimate MiD effects as introduced earlier in the paper.

Overlap Assessment. Figure A7 reports the distribution of propensity scores for municipalities with a positive price shock (unmoderated) and negative price shock (moderated) municipalities after matching without replacement. Although parts of the domain have common support, we note that the distribution of propensity scores is visibly different across the two groups. This casts doubt on the moderator's exogeneity, as, even after matching, the covariates between moderated and unmoderated units differ. That said, the overlap is sufficient such that we can proceed with the use of our MiD estimators in the next subsection.

Robustness to MiD Estimators. The original HiD estimates match the results in the paper: the incumbency disadvantage is of a greater magnitude in places with a negative price shock. As we proceed to the initial MiD estimator, we observe little difference in the estimates. This changes drastically, however, as we apply our (preferred) matching estimator. Here, the estimates flip their sign and become statistically insignificant: we can no longer reject the null that incumbency (dis)advantages are similar across municipalities with and without negative price shocks. This offers further evidence, on top of the distribution of propensity scores, that the conditional independence assumption is unlikely to hold with estimators without covariate adjustment, and that the results in the paper do not capture the intended MiD.

The application of our framework to this case highlights two key issues. First, even where common support between moderated and unmoderated units exists, the moderator may not be assigned randomly or quasi-randomly: additional covariate adjustment may be necessary to move closer to satisfying the conditional independence assumption. Second, the contrast between the HiD and MiD estimates underscores the contribution of our framework. This distinction is also meaningful for policymaking: policies with the intention to reduce incumbency disadvantages by focusing on price stability may prove ineffective if, after all, price shocks are not a causal moderator. A further caveat that adds to the difficulty of capturing the direct moderation

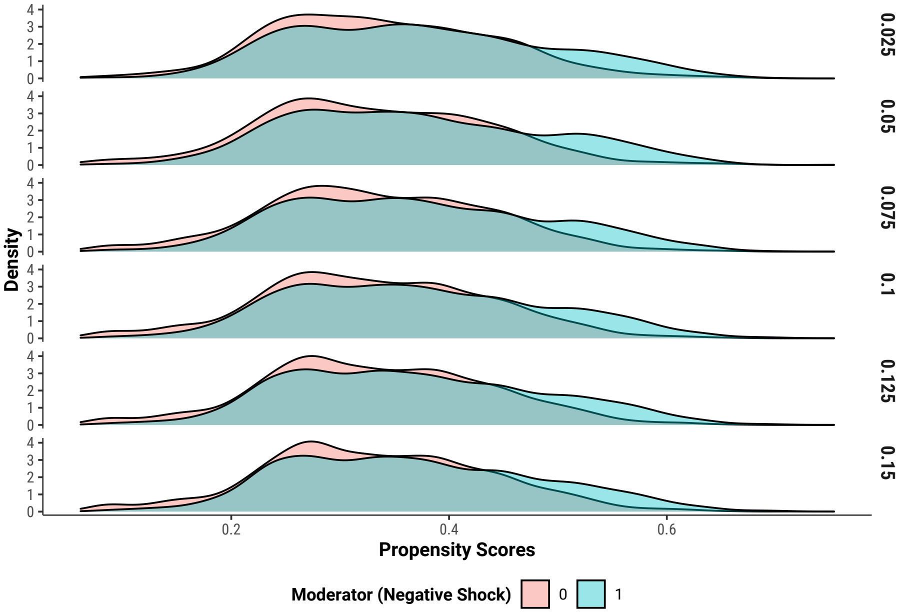

Figure A7: Assessing overlap among observations in Novaes and Schiumerini (2021), by bandwidth. We plot the distribution of propensity scores (measuring how likely an observation is in the moderated group) after matching on the full control set.

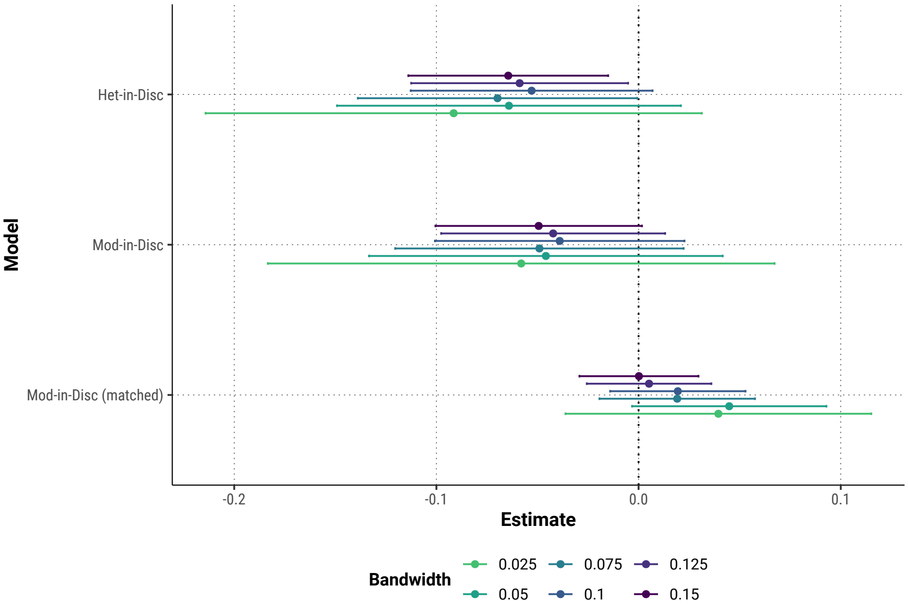

Figure A8: Moderation-in-Discontinuities estimates on the moderation effect of price shocks on incumbency (dis)advantages, following Novaes and Schiumerini (2021).

effect is the interplay of the price shock moderator with covariates such as GDP, rendering evaluation of the true causal mechanism particularly difficult.

E.2 The Effect of General Primaries on Incumbency Advantages (Olson, 2020)

Another application of our MiD framework, following Olson (2020), investigates the difference in incumbency advantages as a result of introducing direct primaries in US House elections. As before, we follow our framework by measuring overlap and fitting our MiD estimator. We caution that the original design makes strong causal claims without necessary formalization, and find that the two groups (moderated and unmoderated) have no overlap, which renders the estimation of the true MiD effect infeasible.

E.2.1 Original Design and Interpretation.

Olson (2020) studies whether the method used to nominate party's candidates in U.S. House elections affects the magnitude of the incumbency advantage. The paper argues that the incumbency advantage grew stronger with the introduction of direct primaries, whereas the effect was smaller in previous electoral contexts in which candidates where chosen by party elites. The study evaluates this argument empirically in the context of the gradual adoption of direct primaries across U.S. states in the early 20th century.

The paper is explicit in its interest in the *causal* effect of the nomination regime on the incumbency advantage, which maps onto our MiD estimand.[16]. A precisely estimated moderation effect can help policymakers and electoral reformers help understand the tradeoffs between different nomination regimes and their implications for accountability and selection of candidates into running for office. Of course, there could be lots of of other electoral characteristics that may correlate with direct primaries: differences in voter registration regimes, the strength of political parties, or the salience and dimensionality of various issues in the political arena. In this instance, the Heterogeneities-in-Discontinuities is less useful to policymakers: it remains unclear what actionable insights reformers and decision-makers are left with if the difference in incumbency advantages is merely correlated with direct primary regimes.

Despite the interest in the causal moderation effect, the original design in Olson (2020) fits an estimator that does not map onto either the MiD or the HiD. The running variable is candidates' margin of victory (or loss), and treatment is assigned to those candidates that win. All terms are interacted with the conditioning variable – a binary indicator for nomination regime. What sets the design apart from the classic HiD is the use of an outcome (whether the candidate is elected in $t + 1$) that is demeaned with respect to unit and year (akin to year and unit FEs). This feature intends to identify the causal effect of the direct primary introduction using a setup that leverages the panel structure of the data (with a staggered adoption of direct primaries across U.S. states), but falls short of a formal discussion of the identification assumptions necessary to do so. An extension of this specification includes linearly added covariates (without interactions). We caution that, mirroring the discussion earlier in the paper, identifying the moderation effect in a setting where the treatment is staggered over time is particularly challenging.

To Olson's credit, the paper reports results from an additional specification in which two additional control variables are interacted with the running variable and treatment: civil service regime type and ballot type. These additional variables move the design closer to the MiD, but also feature attenuated (and statistically insignificant) magnitudes of the conditioning effect.

E.2.2 Applying Our Framework

Control Set. In addition to civil service regime time and ballot type, we add additional covariates (used by the author as simple controls, rather than interactions) to the control set: logged population, per cent urban population, and per cent black population. Lastly, we also include the election year as a variable in the control set, though once more caution that the use of time variables in this setting may be more difficult.

Overlap assessment. Figure A9 demonstrates the balance of propensity scores after matching (without replacement) on the control set, within different bandwidths. We observe that there is little overlap

[16] 'The models strongly suggest that direct primary adoption causes an increase in the incumbency advantage' (p. 15)

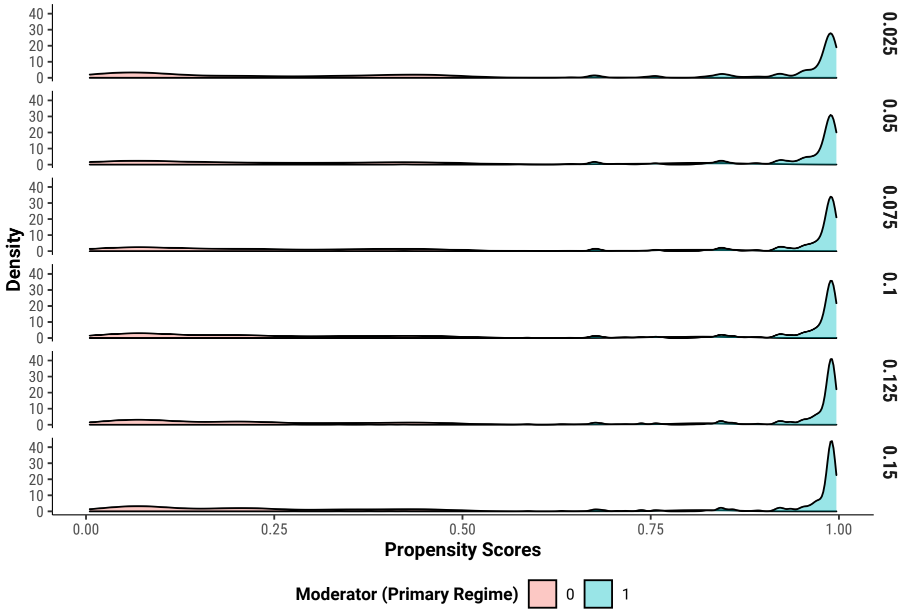

Figure A9: Assessing overlap among observations in Olson (2020), by bandwidth. We plot the distribution of propensity scores (measuring how likely an observation is in the direct primary group) after matching on the full control set (including election year as a continuous covariate).

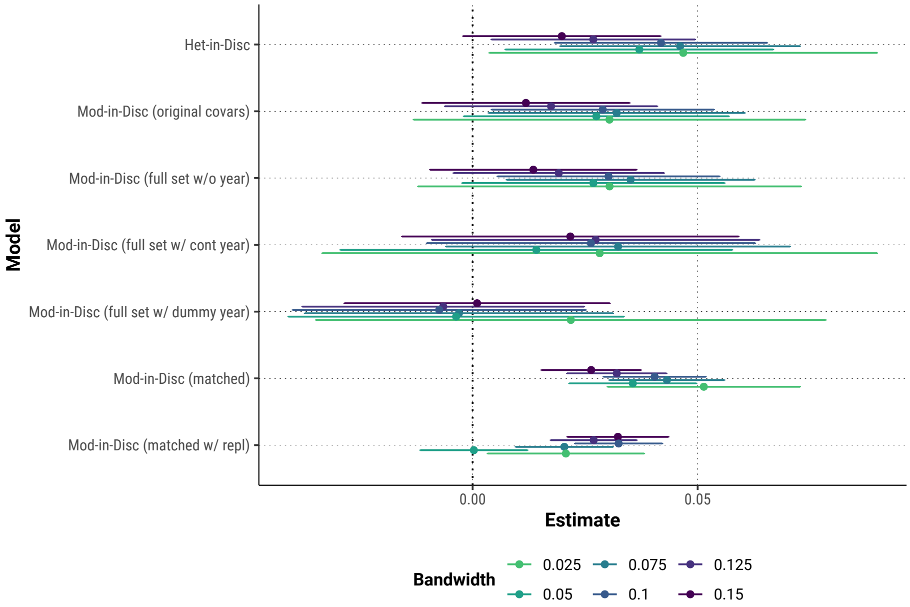

Figure A10: Moderation-in-Discontinuities estimates on the effect of direct primaries on incumbency advantages in the early 20th century U.S., following Olson (2020).

between elections held under the direct primary regime and those without. This alone suggests that a causal interpretation of the HiD is unwarranted. When matching *with* replacement, the overlap improves somewhat, albeit not sufficiently to warrant satisfying the identification assumption.

We caution that the analysis should stop here (as far as a causal estimate of the effect of primary introduction is sought). For the purpose of illustrating the MiD framework further, we nonetheless proceed with presenting our estimates.

Robustness to MiD Estimators. Figure A10 reports the vanilla HiD estimate along with a number of different MiD estimates, evaluated at different bandwidths. First, we emulate the paper's specification and only include the originally interacted covariates in the control set. We note that the estimates drop in magnitude across all bandwidths, and become statistically insigificant except at $h = 0.1$ and $h = 0.075$. We obtain a similar result when we expand the control set to include log population, share of urban population, and the share of Black population (row 3).

We incorporate the election year in two different ways into the control set: as a continuous variable (row 4) and as a dummy indicator for every election year (row 5). In both cases, the confidence intervals grow very wide; in addition, when modeling election years more flexibly, the point estimates also drop close to zero and turn negative. This illustrates how the assumptions we make about time in the MiD framework can have huge impacts on the estimates.

Finally, we also include MiD estimates from our matched datasets. When we match without replacement (row 6), we obtain estimates that are very similar to the original HiD estimator. This is likely an artefact of there being many more treated units than control units: it is simply not possible to obtain close matches, and so the advantage from matching is reduced. When we do allow for matching with replacement (row 7), the estimates attenuate and exhibit a high degree of sensitivity to the bandwidth choice.

Our various MiD estimates point to the following conclusions. First, we highlight the need to assess observations' overlap to justify any interpretation of estimates as moderation effects. Second, the sensitivity of estimates to the inclusion of year dummies shows how precarious the conditional independence assumption is: it is enough for one key variable to be omitted from the control set to distort results. Last, designs that involve panel observations over time with within-unit changes in the moderator are difficult to identify, and can be very sensitive to modeling and identification assumptions. In this application, and even putting common support concerns aside, the high variance in estimates across different specifications leaves us with little confidence that we can identify the true MiD.

References

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*.

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

Casarico, A., Lattanzio, S., and Profeta, P. (2022). Women and local public finance. *European Journal of Political Economy*, 72:102096.

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*.

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.

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*.

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. 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.

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.

Grembi, V., Nannicini, T., and Troiano, U. (2016). Do fiscal rules matter? *American Economic Journal: Applied Economics*, 8(3):1–30.

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.

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

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.

Lassébie, J. (2020). Gender Quotas and the Selection of Local Politicians: Evidence from French Municipal Elections. *European Journal of Political Economy*, 62(101842).

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.

Lindo, J. M., Sanders, N. J., and Oreopoulos, P. (2010). Ability, gender, and performance standards: Evidence from academic probation. *American Economic Journal: Applied Economics*, 2(2):95–117.

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

McCrain, J. and O'Connell, S. D. (2022). Experience, institutions, and candidate emergence: The political career returns to state legislative service.

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.

Novaes, L. M. and Schiumerini, L. (2021). Commodity shocks and incumbency effects. *British Journal of Political Science*, pages 1–20.

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.

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

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.

Wasserman, M. (2018). Gender differences in politician persistence. *The Review of Economics and Statistics*, pages 1–46.

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