**ARTICLE**

# A comprehensive history of regression discontinuity designs: An empirical survey of the last 60 years

**Mauricio Villamizar-Villegas[1]** | **Freddy A. Pinzon-Puerto[2]** |
**Maria Alejandra Ruiz-Sanchez[1]**

[1] Central Bank of Colombia, Bogota, Colombia

[2] Universidad del Rosario, Bogota, Cundinamarca, Colombia

**Correspondence**
Mauricio Villamizar-Villegas, Central Bank of Colombia, Bogota, Colombia.
Email: mvillavi@banrep.gov.co

**Abstract**

In this paper we detail the entire Regression Discontinuity Design (RDD) history, including its origins in the 1960s, and its two main waves of formalization in the 1970s and 2000s, both of which are rarely acknowledged in the literature. Also, we dissect the empirical work into fuzzy and sharp designs and provide some intuition as to why some rule-based criteria produce imperfect compliance. Finally, we break the literature down by economic field, highlighting the main outcomes, treatments, and running variables. Overall, we see some topics in economics gaining importance through time, like the cases of health, finance, crime, environment, and political economy. In particular, we highlight applications in finance as the most novel. Nonetheless, we recognize that the field of education stands out as the uncontested RDD champion through time, with the greatest number of empirical applications.

**KEYWORDS**
empirical survey, fuzzy and sharp designs, regression discontinuity design, RDD formalization

**JEL CLASSIFICATION**
B23, C14, C21, C31, C52

*"…a discontinuity, like a vacuum, is abhorred by nature"* (Hotelling), yet *"fluidity and discontinuity are central to the reality in which we live in"* (Bateson)[1]

## 1 | INTRODUCTION

In this paper we bring together the entire Regression Discontinuity Design (RDD) literature; namely 60 years in the making, and covering 239 empirical papers. Our contribution is three-fold. First, we detail a comprehensive RDD history, including its origins (in the 60s) and its two main waves of formalization (in the 70s and 00s), both of which are rarely acknowledged in the literature. Second, we dissect the literature into fuzzy and sharp designs and provide some intuition as to why some rule-based criteria produce imperfect compliance, i.e. cases in which treated and control observations are only partially enacted. Finally, we break the literature down by economic field, highlighting the main outcomes, treatments, and running variables.

In contrast to practical user guides such as Imbens and Lemieux (2008), Cook and Wong (2008), Lee and Lemieux (2010), DiNardo and Lee (2011), and Cattaneo et al. (2019), we do not provide a step-by-step handbook of implementation procedures. Rather, we provide a comprehensive outlook of the vast variety of existing designs within each field. Thus, RDD practitioners can quickly extract and focus on the strand of literature closest to their investigation and be warned about potential perils and pitfalls. We also provide a general theoretical framework to familiarize readers with RDDs viewed as *ex-post facto* experiments. We finally reference influential studies on some of the most technical issues.

The RDD history begins in the 1960s, when Thistlethwaite and Campbell (1960) used a discontinuity approach to measure the effect of students winning certificates of merit when applying to college. A few years later, Sween and Campbell (1965)'s contribution to interrupted time series paved the way to separately characterizing sharp and fuzzy designs. Surprisingly, the methodology lay practically dormant for almost 40 years, except for the period that we denote as the *first wave* in the 1970s, led by Goldberger (1972), Barnow (1972), Rubin (1974), and Trochim and Spiegelman (1980), among others. This period is largely portrayed by the pairings between randomized controlled trials (RCTs) and approximations to true experiments (quasi-experiments). When both are run in tandem, estimators can be combined and compared in terms of precision (efficiency) and unbiasedness. At this point in time, however, RDDs were either viewed as a special case of Difference-in-Differences (Sween & Campbell, 1965), Matching (Heckman et al., 1999), or Instrumental Variables (Angrist & Krueger, 2001).

At the turn of the new millennium, we document the revival of RDDs with its *second wave* of formalization, led by authors such as Hahn et al. (2001), Porter (2003), and Leuven and Oosterbeek (2004). In particular, we acknowledge the pioneer works of Hahn et al. (1999) and Angrist and Lavy (1999) in formally conducting sharp and fuzzy RDDs, respectively, in the way that they are applied today.

Overall, there are several attractive features that embody RDDs, including the seemingly mild assumptions needed for identification. Further, a well-executed RDD produces similar estimates to those from RCTs made at the same cutoff value, indicating a high internal validity. This has led authors like Lee and Lemieux (2010) to state that RDDs are "*a much closer cousin of randomized experiments than other competing methods.*" Barnow (1972), Trochim and Spiegelman (1980), Cook and Wong (2008), and Chaplin et al. (2018) empirically corroborate this by contrasting cases in which RDDs and purely experimental designs are conducted (with overlapping samples) and

show that, based on different criteria, they yield similar results. External validity of RDDs on the other hand, depend on whether the assumption of homogeneous treatment effects holds.

On the downside, it is not often the case that policy treatments are deterministically triggered, so finding a suitable real-life case study is challenging. Also, a subset of problems that apply to standard linear estimations carry over to RDDs (e.g., misconstruing non-linearities for discontinuities). We also recognize that RDDs are subject to greater sampling variance, and thus have less statistical power than a randomized experiment with equal sample size, sometimes by up to a factor of three (Goldberger, 1972). Finally, being a localized approach, if the design narrows in too close to the threshold, less observations become available. This is why RDD is ultimately an extrapolation-based approach. Notwithstanding, we show that the economic literature is growing ever more receptive to studies employing RDDs.

In the entire literature, we see some topics in economics gaining importance through time, like the cases of finance, health, crime, environment, and political economy. In particular, we highlight applications in finance as the most novel, which, to the extent of our knowledge, no other survey covers. In this area, most of the research has centered on the effects of credit default on variables such as corporate investment, employment, and innovation. In low-income countries, RDDs have been used to study the impact of microcredit programs on output, labor supply, schooling, and even reproductive behavior. Additionally, some work on finance use RDDs in procurement auctions, where neighboring bids reveal a similar valuation of the underlying asset, and thus bidders are locally very similar to each other except in actually winning the auction. Finally, we highlight a novel wave of RDD studies related to applications on Fintech (e.g., access to Fintech loans). Nonetheless, we recognize that the field of education stands out as the uncontested RDD champion through time, with a total number of 62 empirical studies. The area of finance is the runner-up, with a total of 42 empirical studies and the area of labor is the second runner-up, with 38 empirical studies.

Our RDD inventory was compiled through a Web-scrapping search across different sources, including RePEc, Scopus, Mendeley, and Google Scholar. We searched for studies containing *Regression + Discontinuity + Design* (plural and singular) either in the title or abstract. To avoid leaving some papers out of our analysis, we also conducted robustness searches, for example, studies with JEL classification categories pertaining to the letter "C" (in particular C14, C21, C31, and C52). After manually revising each search result and keeping only the most updated or published version of each paper, we gathered a total of 239 studies, all of which are presented in our Online Appendix. It reports each paper's corresponding field, outcome variable, treatment, running variable, cutoff value (and unit of measure), whether it employs a fuzzy or sharp design, and key findings. Finally, for each paper we obtained the number of citations and *Web of Science-ISI* impact factor (if published). Those with highest scores (most influential) are largely analyzed in Sections 3 and 4.

We believe that RDD practitioners of all fields can benefit from our survey. For one, it complements existing surveys and user guides such as Lee and Lemieux (2010), by including at least a decade's worth of new findings. Hence, while Lee and Lemieux cover approximately 80 studies up until 2009, we cover 239 studies up until 2021, a significant nearly three-fold increase in sample size.[2] Also, we build on some of the extensions proposed by Lee and Lemieux. In particular, the authors state that "*there are other departures that could be practically relevant but not as well understood. For example, even if there is perfect compliance of the discontinuous rule, it may be that the researcher does not directly observe the assignment variable, but instead possesses a slightly noisy measure of the variable. Understanding the effects of this kind of measurement error could further expand the applicability of RD*" (p. 351). We make emphasis on this point and more generally on

cases where treatment is partially enacted at the threshold. We also provide numerous examples of these fuzzy designs in Section 4, that range from financial options not being exercised, to eligible beneficiaries of government programs who are unwilling to take part. Finally, we provide some intuition as to why some rule-based criteria produce perfect and imperfect compliance even when using the same running variable.

## 2 | THEORETICAL UNDERPINNINGS OF RDDS

This section sets the stage for the basic understanding of Regression Discontinuity Designs (**RDDs**). While it intends to familiarize readers with its general framework, it does not provide a thorough examination of some of its more technical aspects. In those cases, however, we do direct readers to the most recent studies that have contributed to its overall formalization. For ease in readability, in Table 9 we provide a glossary of the RDD terminology.

### 2.1 | RDD viewed as an *ex-post facto* experiment

Essentially, RDDs are characterized by a deterministic rule that assigns treatment in a discontinuous fashion. For example, in Thistlethwaite and Campbell (1960), students are awarded a certificate of merit based on the CEEB Scholarship Qualifying Test. Exposure to treatment (i.e., receiving the certificate of merit) is determined by scoring above a given grade, creating a discontinuity of treatment right at the cutting score. Hence, the main underlying idea of the design is for students scoring just below the threshold (control group) to serve as valid counterfactuals to students who barely crossed the threshold (treatment group) had they not received the award.

More formally, in the standard *Sharp* RDD setup, the assignment of treatment, $D_i$, is completely determined by a cutoff-rule based on an observable (and continuous) **running variable**, $X_i$, as follows:

$$D_i = \mathbf{1}\{X_i \geq x_0\} \tag{1}$$

where $\mathbf{1}$ denotes an indicator function and $x_0$ is the threshold below which treatment is denied. The discontinuity arises because no matter how close $X_i$ gets to the cutoff value from above or below (as in the case of test scores), the treatment, or lack of treatment, is unchanged. Intuitively, the rule creates a natural experiment when in close vicinity of $x_0$. If treatment has an effect, then it should be measured by comparing the conditional mean of an outcome variable (or vector of outcome variables) at the limit on either side of the discontinuity point:

$$\text{Average Treatment Effect} = E(Y_{1i} - Y_{0i} \mid X_i = x_0)$$

$$= E(Y_{1i} \mid X_i = x_0) - E(Y_{0i} \mid X_i = x_0)$$

$$= \lim_{\epsilon \downarrow 0} E(Y_i \mid X_i = x_0 + \epsilon) - \lim_{\epsilon \uparrow 0} E(Y_i \mid X_i = x_0 + \epsilon) \tag{2}$$

where, for each individual $i$, there exists a pair of potential outcomes: $Y_{1i}$ if exposed to treatment, and $Y_{0i}$ if not exposed. The final equality holds as long as the conditional distributions of potential

outcomes, $\Pr(Y_{1i} \leq y \mid X_i = x)$ and $\Pr(Y_{0i} \leq y \mid X_i = x)$, are continuous at $X_i = x_0$. Namely, as will be explained in Section 2.3, this requires individuals to be incapable of precisely controlling the running variable.

To better conceptualize the RDD methodology as an *ex-post facto* experiment, consider an example found in Kuersteiner et al. (2018), who initially propose the following linear regression model:

$$y_i = \alpha + \beta D_i + \gamma_i. \tag{3}$$

In the context of merit-based awards such as the one described in Thistlethwaite and Campbell (1960), it is evident that an Ordinary Least Squares (OLS) estimate of $\beta$ does not precisely capture the effect of treatment. That is, students with very good test scores (far from the cutoff) are likely to also embody characteristics of high effort and ability. Thus, when evaluating the effect of the scholarship award on future income, these other characteristics would mask the true effect of treatment, leading to an upward bias. More formally, the bias can be computed as the conditional mean of the error term with and without treatment, $E[\gamma_i | D_i = 1] - E[\gamma_i | D_i = 0]$, which would most likely be positive in this context.

In contrast, consider the same linear model but under a localized analysis around the cutoff score. Any endogenous relationship is now broken down by the fact that small variations in the running variable ($X_i$), which lead to small variations in $\gamma_i$, generate a discontinuous jump in $D_i$. It is precisely this variation that an RDD exploits in order to conduct causal inference. Formally,

$$\lim_{\epsilon \downarrow 0} E[y_i | X_i = x_0 + \epsilon] - \lim_{\epsilon \uparrow 0} E[y_i | X_i = x_0 + \epsilon]$$

$$= \left( \alpha + \beta + \lim_{\epsilon \downarrow 0} E[\gamma_i | X_i = x_0 + \epsilon] \right) - \left( \alpha + \lim_{\epsilon \uparrow 0} E[\gamma_i | X_i = x_0 + \epsilon] \right) = \beta. \tag{4}$$

As noted in Equation (2), the RDD approach relies on the somewhat weak assumption that unobservable factors vary smoothly around the cutoff. This is also exemplified in equation (4) where the two conditional means of the error term (with and without treatment) cancel out at the limit, when $X_i = x_0$. Further reading on this continuity assumption is presented in Hahn et al. (2001), Porter (2003), Imbens and Lemieux (2008), and Lee (2008). It allows for the treatment (e.g., certificate of merit) to become uncoupled from unobservable factors when narrowing locally at the threshold.

Imbens and Lemieux (2008) point out that local regressions improve over simple comparison of means around the discontinuity point as well as over standard kernel regressions. In turn, Kuersteiner et al. (2016) provide a useful representation of RDDs applied to a time series setting and with the use of local projections, as in Jordà (2005). The resulting representation, based on Hahn et al. (2001), but with a cutoff-rule adapted to a time series setting, where $D_t = \mathbf{1}\{X_t \geq x_0\}$, is presented as follows:

$$\left( \hat{a}, \hat{b}, \hat{\gamma}, \hat{\theta} \right) = \underset{a,b,\gamma,\theta}{\arg \min} \sum_{j=1}^{J} \sum_{t=2}^{T-J} \left( y_{t+j} - a_j - b_j(X_t - x_0) - \theta_j D_t - \gamma_j(X_t - x_0)D_t \right)^2 K\left( \frac{X_t - x_0}{h} \right) \tag{5}$$

where $\theta = (\theta_1, \dots, \theta_J)'$ are the impulse-response coefficients that capture the impact of treatment on outcome variables $\}\}j''$ periods after treatment, the term $K(\cdot)$ represents a kernel function (where $h$ is its bandwidth parameter), and $b_j$, and $\gamma_j$ are polynomial functions of the running

variable. The inclusion of the term $\gamma_j(\cdot)D_t$ allows for different specifications of how the running variable affects the outcome, at either side of the cutoff.

An attractive feature of RDDs is that, by design, the Conditional Independence Assumption (CIA) is satisfied. For readers unfamiliar with this assumption, the CIA states that conditional on an informative history, policies are independent of potential outcomes, or as good as randomly assigned. This allows for the foundation based on which *"regressions can also be used to approximate experiments in the absence of random assignment"* (Angrist and Pischke (2009), p. 18). Specifically, the CIA can be formulated as:

$$Y_{kt} \perp D_t \mid X_t \qquad \text{for } k = 0, 1. \tag{6}$$

In the context of RDDs, the running variable carries all the information needed to construct the policy variable, so the term $D_t \mid X_t$ is purged from *all* information (recall that $D_t$ is a deterministic function of $X_t$). This is the reason why the CIA is trivially met. The drawback, however, lies in the impossibility of observing both *treated* and *control* observations for a given value of $x_t$, something that in the literature is referred to as the overlap assumption. Hence, in any RDD approach, the continuity assumption of potential outcomes plays an essential role, namely, to compensate for the failure of the overlap condition. And, even though the continuity assumption cannot be purely tested, it does have some testable implications which are discussed in Section 2.3.

We next turn our attention to cases in which treatment is partially enacted at the threshold. In other words, where there is imperfect compliance. Examples of these fuzzy designs range from financial options not being exercised, to eligible beneficiaries of government programs who are unwilling to take part. In Section 4, we detail various examples covered in the literature. In these cases, a valid design is still possible as long as there is a discontinuous jump in the probability of being treated, even though the running variable does not perfectly predict the treatment status. This necessary condition is exemplified as follows:

$$\lim_{\epsilon \downarrow 0} Pr(D_i = 1 \mid X_i = x_0 + \epsilon) \neq \lim_{\epsilon \uparrow 0} Pr(D_i = 1 \mid X_i = x_0 + \epsilon). \tag{7}$$

For the most part, studies that conduct a fuzzy design carry out an Instrumental Variable (IV) approach, by instrumenting the observed treatment status (i.e., whether an individual was actually treated or not) with both the running variable and the treatment dictated by the cutoff rule. As documented in Lee and Lemieux (2010), the treatment effect can then be computed as the ratio between the jump in the outcome variable and the share of compliant observations (those that are triggered by the rule and receive treatment), as follows:

$$\frac{\lim_{\epsilon \downarrow 0} E[Y_i | X_i = x_0 + \epsilon] - \lim_{\epsilon \uparrow 0} E[Y_i | X_i = x_0 + \epsilon]}{\lim_{\epsilon \downarrow 0} E[D_i | X_i = x_0 + \epsilon] - \lim_{\epsilon \uparrow 0} E[D_i | X_i = x_0 + \epsilon]}. \tag{8}$$

Notice that Equation (8) represents a simple Wald-type IV estimator. That is, there is a close analogy between how the treatment effect is defined in the fuzzy RDD and the often-referred Wald formulation of the treatment effect in an IV approach. Hahn et al. (2001) were the first to show this relevant link and to recommend estimating the treatment effect using a two-stage least squares procedure in this setting (the authors also offered an interpretation of the Wald estimator as an RD estimator).

## 2.2 | Causal inference and external and internal validity

At face value, an RDD estimand, because of its localized nature, is supposed to be informative only at the cutoff value. In the example presented in Thistlethwaite and Campbell (1960) this would mean that the effect of winning a certificate of merit applies only to students with test scores in the vicinity of the passing grade and uninformative anywhere else.

However, we take a more optimistic view. As noted in Lee and Lemieux (2010), the RDD estimand can be seen as a weighted average treatment effect, where the weights are directly proportional to the ex-ante probability that an individual's realization of her running variable will fall within the immediate neighborhood of the threshold. Hence, in principle, *all* individuals (or measurable units) could weigh-in. One example is when using a kernel function with a sufficiently large bandwidth (using information far from the discontinuity point). Ultimately, the similarity of weights brings the RDD estimand closer to the overall average treatment effect. So, under the (rather strong) assumption of homogeneous treatment effects, RDDs have a high external validity.

In the related literature, Chaplin et al. (2018) examine the external validity of RDDs by evaluating how the size of the *RD bias* varies across 15 studies with different settings, where the *RD bias* is defined as the difference between a local RD estimate at the cutoff and an RCT estimate made at the same point. The authors find that when the study-specific estimates are shrunken to capitalize on the other studies' information, all estimates fall within 0.07 standard deviations of their RCT counterparts, indicating a high external validity (with unshrunken estimates, the mean bias is essentially zero). In turn, Bertanha and Imbens (2020) explore the external validity in fuzzy RDDs. Specifically, the authors test for the equality of the distributions of potential outcomes for treated compliers and always-takers, and for non-treated compliers and never-takers. They argue that if both equalities hold, then it more likely to extrapolate the average effect for compliers to other sub-populations (higher external validity).

In practice, these reasons are why numerous studies normalize (or at least de-mean) the running variable in order to pool observations together, regardless of whether they belong to different cutoffs. Unfortunately, a major caveat is that it is only possible to observe one realization of the running variable for each individual, so it is impossible to empirically test for the similarity of weights *across* individuals. To do so, the design would require information about each person's ex-ante probability distribution of the running variable. Nonetheless, this "limitation" is shared with many other methods. For instance, in an IV approach, the effect is, in principle, only applicable to the sub-population whose treatment is affected by the instrument.

Regarding internal validity, RDDs have more in common with RCTs than with practically any other method, including Matching, IVs, and OLS (Lee & Lemieux, 2010). And, while some authors consider RDDs to be a special case of selection-on-observables (Heckman et al., 1999), we believe that it differs in one key aspect, namely that *"one need not assume that RD isolates treatment variation that is as good as randomly assigned, instead randomized variation is a consequence of agents' inability to precisely control the assignment variable near the known cutoff"* (Lee & Card, 2008). As stated in the previous section, this is why the Conditional Independence Assumption (CIA) is trivially met. In all other selection-on-observables methods, this condition needs to be assumed. Analogously, in IVs the instrument is assumed to be exogenously generated in order to meet the exclusion restriction.

The literature on the internal validity of RDDs is more abundant and, among many, we document the works of Barnow (1972), Trochim and Spiegelman (1980), Cook and Wong (2008), and Chaplin et al. (2018) which compare cases in which RDDs and purely experimental designs are

conducted (with overlapping samples) and show that, based on different criteria, they yield similar results. With this in mind, in Section 3.2 we highlight that, while RDDs produce unbiased estimates, they are subject to greater sampling variance, and on average RCTs are 2.75 times more efficient than RDDs, due to the correlation between the assignment variable and the cutoff (Goldberger, 1972).

## 2.3 | Specification testing: Perils and pitfalls

In order to have a locally randomized experiment, market participants cannot perfectly control the running variable near the cutoff (Lee, 2008). To illustrate, McCrary (2008) provides an example in which workers are only eligible for a training program as long as their income in a given period is at or below a hypothetical value of $14. He shows that manipulation of the running variable yields too few (and likely different) workers above the threshold, and too many workers just below. In principle, this assumption is fundamentally untestable given that we observe only one realization of $X$ for a given individual ($X_i$) or time period ($X_t$).

Nonetheless, McCrary proposes a test that has become standard in the RDD literature: to analyze the bunching of observations in close proximity of the threshold. In other words, the test estimates the density function of the running variable at either side of the cutoff. Rejecting the null hence constitutes as evidence of precise sorting (i.e., manipulation) or self-selection around the cutoff, a clear warning signal that can compromise the validity of the design.[3] However, some promising work has been conducted when in the presence of heap-induced bias. For example, Barreca et al. (2016) propose a "donut-RD" approach that estimates treatment after systematically dropping observations in close vicinity of the threshold. As an alternative, the authors also propose an estimation that allows separate intercepts and trends for the heaped data.

In addition, the estimated effect must be attributed solely to the treatment being considered. To avoid type-I errors (i.e., a false positive), Hahn et al. (2001) argue that all other factors must be evolving smoothly with respect to the running variable. Put differently, the distribution of baseline covariates should not change discontinuously at the threshold. In the literature, designs based on population thresholds are generally subject to this pitfall, since the same population threshold is often used to determine multiple policies (Eggers et al., 2018). As a potential solution, Grembi et al. (2016) propose a *difference-in-discontinuities*, where the treatment of interest can be identified in a specific period or setting, different than other periods where other policies are run in tandem.[4] Alternatively, to evaluate whether results are exclusively due to treatment, some authors propose the analysis of "placebo" jumps, which replaces the true cutoff with an artificial threshold value in the running variable (Imbens, 2004). These fake-cutoff tests are described in Cattaneo et al. (2019) and De Magalhaes et al. (2020) as indirect tests of the continuity assumption, arguing that jumps in the conditional expectation of potential outcomes at arbitrary points cast doubts on inference and on the overall reliability of the estimates.

If the design achieves local random variation, then there is no strict need to include control variables since treatment is, by design, independent of potential outcomes. Nonetheless, some authors place emphasis on the inclusion of covariates in order to increase accuracy in RDD estimates (by reducing the sampling variability in the estimator). However, if baseline covariates are not balanced (Black et al., 2005, Calonico et al., 2019, and Imbens & Lemieux, 2008) or if there is evidence of manipulative sorting in the running variable (Eggers et al. (2018)), then control variables might provide some oxygen life support to the design. To this effect, Frölich and Huber (2018) propose a non-parametric method to account for observed covariates and show that it can

reduce the asymptotic variance as well as control for discontinuities in the covariate distribution. In turn, Linden and Adams (2012) proposes a parametric solution, namely, to apply Propensity Scoring techniques to adjust for observed differences in baseline characteristics. This involves multiplying the *inverse probability of treatment weights* by the kernel weight and using this new weight in the regression. We note that in some cases, a researcher needs to redefine the parameter of interest.

Furthermore, a key concern in RDD analysis has to do with non-linearities, often misconstrued as discontinuities. In this sense, a researcher generally decides over implementing a parametric or non-parametric approach when estimating the effects of treatment. As stated in Lee and Lemieux (2010) "*If the underlying conditional expectation is not linear, the linear specification will provide a close approximation over a limited range of values of X (small bandwidth), but an increasingly bad approximation over a larger range of values of X (larger bandwidth)*". Note that if the methodology narrows in too close to the threshold, less observations become available. This is why RDD is fundamentally an extrapolation-based approach. A useful discussion on the choice of estimator is found in Hahn et al. (2001), Porter (2003), and Imbens and Lemieux (2008). Everything considered, parametric and non-parametric estimations should be seen as complementary and ideally report similar findings.

Fortunately, a very recent (and growing) strand of literature has centered on the choice of bandwidth, which is critical since it determines which observations are included in the analysis. Given that a "correct" bandwidth parameter can only be unveiled when knowing the true relationship between the outcome and running variables, several sensitivity tests have been proposed. According to Lee and Lemieux (2010), there are typically two approaches for choosing bandwidths. The first is based on a cross-validation or asymptotic mean squared error minimization procedure, which optimally balances the inherent tradeoff between bias and variance. Consequently, choosing a small (large) bandwidth around the threshold will diminish (increase) the misspecification error, thus reducing (increasing) bias, but with a larger (smaller) variance, so essentially with less (more) precision. For further details, we refer reader to the works of Ludwig and Miller (2007) and Imbens and Lemieux (2008).

The second approach consists of characterizing the optimal bandwidth in terms of the unknown joint distribution of all variables. The relevant components of this distribution can be estimated and then inserted into the optimal bandwidth function (Fan & Gijbels, 1996). Intuitively, a rule-of-thumb bandwidth is estimated over the whole sample, and then used to estimate the optimal bandwidth right at the cutoff.

Currently, the most widely used method is the one described in Imbens and Kalyanaraman (2012), which falls under the first approach (optimal bandwidth under a squared error loss). More recently, Calonico et al. (2020) have proposed a bias-correction of this procedure, where the extra variability induced by the bias-correcting term is taken into account. Specifically, the authors first re-center the usual t-statistic with an estimate of the leading bias. Then, to correct for poor finite-sample performance, the authors rescale the t-statistic with a standard error formula that accounts for the additional variability introduced by the estimated bias.

Finally, we note that the treatment effect in RDDs is not only sensitive to the choice of bandwidth, but also to the choice of polynomial order. In one extreme, a researcher can choose to include the entire sample (with a sufficiently large bandwidth) and focus on the global polynomials of the running variable, so as to capture its association with the outcome variable. In the other extreme, any polynomial function, with a sufficiently small bandwidth, starts to look linear. In all intermediate cases, *local* polynomials are usually employed with the purpose of relaxing the linearity assumption and to assess robustness of the results. A word of caution, however, is found

in Gelman and Imbens (2019) who give three reasons why high-order polynomials should not be used. These reasons consist of (i) noisy estimates, (ii) sensitivity to the degree of the polynomial, and (iii) poor coverage of confidence intervals.

## 2.4 | RDD extensions

Several extensions to the standard RDD setting have also been considered in the recent literature. One example is the full extension of RDDs to time-series or panel settings. This is particularly useful when the running variable carries some built inertia and, as a consequence, control episodes get eventually mixed with treated observations as the time horizon expands. Hence, the Stable Unit Treatment Value Assumption—SUTVA—described in Rubin (1974) does not hold. Also, agents can incorporate the possibility of treatment even before the rule is actually triggered. We refer readers to the work Hausman and Rapson (2018) for a description of potential pitfalls affecting RDDs in time, to Kuersteiner et al. (2016) for a technical application of these issues in the context of exchange rate options in a time-series setting, and to Perez-Reyna and Villamizar-Villegas (2019) for a panel setting application in the context of foreign bond holdings.

Additionally, a growing strand of literature has focused on discontinuities found in geographical borders. A pioneer study in this area is Card and Krueger (1994) who compare food establishments in the border between two US states: New Jersey (treatment) and Pennsylvania (control), to examine the effect of raising the minimum wage on employment. More recent studies include Dell (2010) who studies the effects of forced mining labor in Peru and Bolivia, and Spenkuch and Toniatti (2018) who study regulations on political advertising across different counties. In essence, the main challenges of geographical designs, as stated in Keele and Titiunik (2015), include the possibility of (i) spatial correlation, (ii) multiple discontinuities at boundaries, and (iii) multiple identification assumptions for the different measures of distance from the boundary. The reason why these challenges arise is due to the fact that geographical borders are seldom set randomly. To correct for these, Keele and Titiunik show that spatial RDDs are equivalent to standard RDDs with the inclusion of two simultaneous running variables: latitude and longitude coordinates.

An application that has gained momentum in the recent RDD literature deals with heterogeneous treatment effects, that is, cases in which the cutoff varies for each unit in the sample. To date, most studies normalize the running variable to produce a pooled RD treatment effect. However, Cattaneo et al. (2016) point out that this common practice does not fully exploit all the available information. Namely, the authors show that the pooled parameter can be interpreted as "*the weighted average across cutoffs of the local average treatment effects for all units facing each particular cutoff value*" (p. 2). Hence, a multi-cutoff estimate that captures treatment effects at every cutoff value will most likely differ from the pooled estimate (except under specific assumptions).[5]

Some studies have tried to tackle the identification challenge posed by measurement error in RDDs. For instance, Dong (2015) stresses that many empirical applications use rounded values of the running variable which are hence discrete (e.g., age in years or birth weight in ounces). The previous leads to inconsistent estimates of treatment effects, even when the true functional form relating the outcome and the running variable is known and correctly specified. As a solution, Dong presents an approach that uses information regarding the distribution of the rounding errors. In turn, Davezies and Le Barbanchon (2017) propose a non-parametric estimator of the treatment effect when the difference between the true running variable and its noisy version is observed in an auxiliary sample of treated individuals. Similarly, Bartalotti et al. (2021) introduce

a procedure that uses auxiliary data to correct the bias induced by group-specific measurement error.[6]

A common concern in empirical applications of RDDs is the issue of sample selection. Intuitively, the standard RDD identification relies on the comparability of observations neighboring the threshold. As such, differential sample selection around the RD threshold may threaten such comparability. Dong (2019) extends the standard RDD to allow for sample selection around the threshold, by providing a nonparametric identification of the "extensive" and "intensive" margin effects. While the former evaluates the treatment effect on the sample selection probability, the latter evaluates the treatment effect on the observed outcome, conditional on being selected into the sample. Based on these identification results and principal stratification, the author constructs sharp bounds for the treatment effects among a well-defined subgroup: those always participating compliers.

Finally, promising RDD work has been conducted in cases where: the running variable is discrete (Lee & Card, 2008), the discontinuity point is unobserved (Porter & Yu, 2015), kink regression design (Chiang et al., 2021 and Nielsen et al., 2010 ), and multiple running variables are employed (Foote et al., 2015). Also, when using quantile treatment effects (Frölich & Melly, 2008), likelihood functions (Otsu et al., 2013), and dynamic treatment effects (Cellini et al., 2010 and Chiang et al., 2019). We believe that future work on the robustness of fuzzy designs' inference will soon largely develop (Feir et al., 2016).

## 3 | A FUZZY HISTORY THROUGH TIME

In the year 1960, Regression Discontinuity was first implemented to measure the effect of 5126 students winning certificates of merit compared to 2848 students who simply received letters of recommendation when applying to college (Thistlethwaite & Campbell, 1960).

Surprisingly, the methodology lay practically dormant for almost 40 years, except for the period that we denote as the first wave of RDD formalization in the 1970s, led by Goldberger (1972), Barnow (1972), Rubin (1974), and Trochim and Spiegelman (1980), among others. This period is largely characterized by the pairings between randomized controlled trials and approximations to true experiments. When both are run in tandem, estimators can be combined and compared in terms of precision (efficiency) and unbiasedness.

At the turn of the new millennium, we document the revival of RDDs with its second formalization wave, led by authors such as Hahn et al. (1999), Angrist and Lavy (1999), Hahn et al. (2001), Porter (2003), Leuven and Oosterbeek (2004), McCrary (2008), and Imbens and Kalyanaraman (2012).

### 3.1 | RDD origins

Campbell and Stanley (1963) are probably pioneers in formalizing RDD procedures, by examining the validity of experimental designs against common threats to causal inference. However, Sween and Campbell (1965)'s contribution to interrupted time series allowed for a more in-depth examination of RDDs and even paved the way for Campbell (1969) to separately characterize sharp and fuzzy designs. That is, to differentiate between cases in which treatment is fully (*i.e., sharp*) or partially (*i.e., fuzzy*) enacted at the threshold. More formally, Sween and Campbell (1965) propose three tests to disentangle treatment from purely random variation:
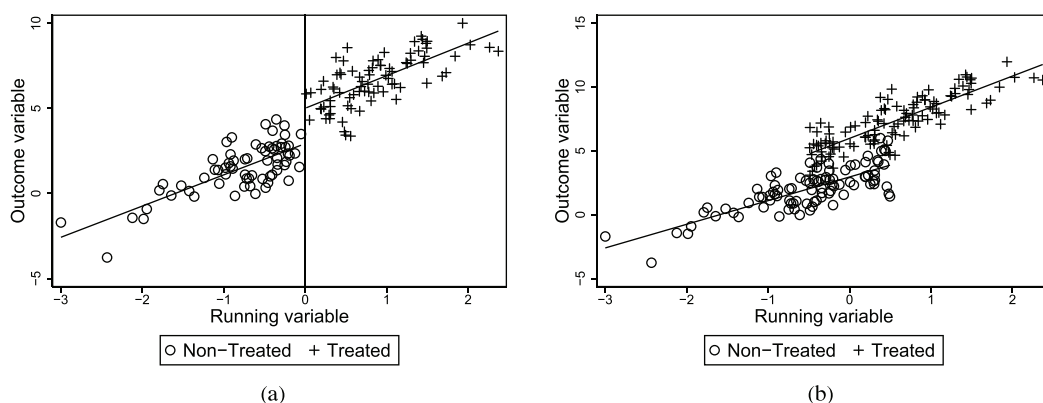
**FIGURE 1**  Computer-generated RDD data taken from Campbell (1969)

1. **Mood test**: T-test significance of the first post-treatment observation from a value predicted by a linear fit based on pre-treatment observations,
2. **Walker-Lev tests**: A set of tests to determine pre and post-existing differences (regression slopes) between treated and control groups, and
3. **Double Extrapolation technique**: Significance test of the difference between two separate regression estimates for a given outcome, for the observation midway between the last pre-treatment point and the first post-treatment point.

While the Mood and Walker-Lev tests share similar features to event studies and difference-in-difference methodologies, respectively, the double extrapolation technique narrows in on the vicinity of treatment, and as such constitutes one of the first attempts to formally obtain localized treatment effects.

Interestingly, Campbell explains why the national merit scholarships were unfit to use a sharp design of actual college fellowships, as ideally depicted in Figure 1a. According to the author, pooling rankings of different groups of students across evaluating committees "*provides one cutting point on an unspecified pooled decision base but fails to provide analogous potential cutting points above and below*", leading to a fuzzy design, exemplified by Figure 1b, in which the CEEB Scholarship Qualifying Test does not perfectly predict treatment. At the time, and without any formal fuzzy toolkits at hand, Campbell makes the following recommendation: "*If political patronage necessitates some decision inconsistent with a sharp cutoff, record these cases under the heading 'qualitative decision rule' and* **keep them out of your experimental analysis**".

## 3.2 | First formalization wave

During the 1970s, a handful of authors sought ways to implement RDDs applied to a variety of scenarios. In most cases, they focused on bias reduction techniques based on the eligibility criteria to receive treatment. This small group of RDD advocates, which favored discontinuity designs as a preferred evaluation approach, are part of what we denote *the first formalization wave*.

We first document Barnow (1972), who evaluates education programs when group selection is not random (or similarly, in the absence of pre-treatment information). Building on some previous work by Campbell and Erlebacher (1970), Barnow presents four models geared to

demonstrate that biases can be avoided if randomization procedures or quasi-experimental techniques are used, including RDDs.

Specifically, the first model considers the case where treatment and control groups are selected from two different samples, with different sample means. In the model, $x^*$ is defined as the true ability. The pre and posttest scores are given by $x_i = x_i^* + u_i$ and $y_i = x_i^* + v_i$, respectively, both of which are erroneous measures of true ability. Namely, if a child gets a score higher than her true pretest ability ($x_i \geq x_i^*$), then there is no prior knowledge about whether she will get a score higher or lower than her true posttest ability. Hence, scores are unbiased but do not reflect true ability. In the second model, gain scores (defined as $y_i - x_i$) are regressed on the treatment dummy. While unbiased, estimates are invalidated under heteroskedasticity or non-zero growth rates of ability over time. The third model includes socioeconomic information so that group assignment is random within each socioeconomic group. Finally, the fourth model considers a discontinuous regression model where children are assigned treatment based on their pretest score. As stated in Barnow (1972), the latter is the only valid case.

In a practical example, Goldberger (1972) analyzes compensatory educational programs that had been criticized because students were not randomly selected (students with lower capacity were selected in the treatment group and students with higher capacity in the control group). Thus, even if the program had no effect on any outcome variable, at face value it reported a negative impact. To illustrate this point more formally, consider the following equation:

$$y_i = \beta_0 + \beta_1 D_i + \beta_2 x_i + \eta_i \tag{9}$$

where $D_i$ is a dummy variable denoting whether the individual received treatment, and the pre and posttest are again defined as erroneous measures of true skill: $x_i = x_i^* + u_i$, $y_i = x_i^* + v_i$. Under the assumption that treatment has a **null effect**, Goldberger assesses two particular cases, namely when:

1. individuals are assigned treatment whenever their true skills lie *below* the true mean skill, $x_i^* < \bar{x}_i^*$
2. individuals are assigned treatment whenever their pretest scores lie *below* the pretest mean, $x_i < \bar{x}_i$

For ease in notation, variables are centered at zero so that $\bar{x}_i^* = \bar{x}_i = 0$. In the first case, and given that we only observe $x_i$, then the regression result attributes spurious effects to treatment in the posttest. Put differently, $\beta_1 \neq 0$ denotes the resulting bias. The source of this bias lies in the imperfect selection procedure that assigns a few low-skilled (high-skilled) individuals to the control (treatment) group. As observed in Figure 2a, differences in the true ability of these cases (those in the control group with negative pretest scores and those in the treatment group with positive pretest scores) are carried over to differences in posttest scores. As a result, the coefficient $\beta_1$, which is captured by the differences in group intercepts, reports a negative bias.

In the second case, note that the control group can include low-capacity individuals with unusually high scores in the pretest, and vice-versa. However, since $D_i$ is completely determined by the pretest score, it does not contain additional information about $x_i^*$ (i.e., the only explanatory power that $D_i$ has on $y_i$ is through $x_i$). As shown in Figure 2b, the bias is now eliminated and $\beta_1 = 0$.

Finally, Goldberger finds that the bias-free RDD estimate (where treatment assignment is based on the pretest score) produces the same unbiased estimate as the case of purely random
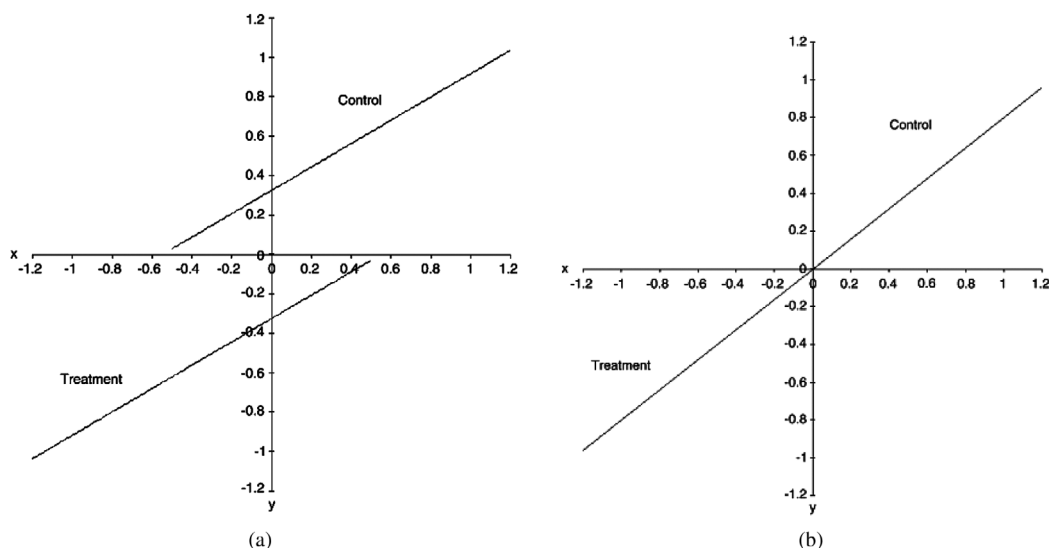
**FIGURE 2**  Selection procedure (pre and posttest scores in x and y-axis); Goldberger (1972)

selection. However, an RDD is subject to greater sampling variance, which means that random selection provides a more efficient design. The magnitude of this efficiency loss is reported as follows: a random sample of size 100 is as good as a selected pretest sample of size 275. Hence, the controlled experiment is 2.75 times more efficient than RDDs, due to the correlation between the assignment variable and the cutoff.

We next turn our attention to Trochim and Spiegelman (1980) who propose an analytical solution known as the "*relative assignment variable approach*" for the problem of pretest-posttest selection bias. In essence, the relative assignment variable is simply the estimated probability of treatment, conditional on the pretest. It bears the same meaning today as a *propensity score*, commonly used in matching techniques. More formally, if we consider the specification in Equation (9), then the validity of the treatment effect, $\beta_1$, will ultimately depend on the estimand of $E[D_i|x_i]$.[7]

Trochim and Spiegelman establish three methods for estimating the probability of treatment.[8] The first is called "*assignment percentage*" where individuals are sorted by their pretest values and divided into equal size intervals. Then, the percent of cases assigned to the program is calculated for each interval, and those values are assigned to individuals. The second method is called "*nearest neighbor moving average*" and essentially computes the moving average of $D_i$'s for the close neighbors of $x_i$, including $x_i$.[9] The last method uses a maximum likelihood Probit model under which the probability of treatment is equal to the estimated cumulative normal distribution, that is, $Pr[D_i = 1|x_i] = \Phi(z_i'\hat{\beta})$.

To exemplify, Trochim and Spiegelman present four model designs with different assignment strategies. The first is a randomized experiment, where $E[D_i|x_i]$ is constant for any pretest value and can be described by a horizontal line relative to $x_i$, as shown in Figure 3a. While this case does not require a pretest, it is included in Equation (9) to increase the statistical power of the program effect estimate. The second model is a sharp design where treatment is based on scoring above a cutoff value. Hence, $E[D_i|x_i] = 0$ for individuals with a pretest value lower than the cutoff and $E[D_i|x_i] = 1$ for those with a pretest value greater or equal than the cutoff, as shown in Figure 3b. Finally, the third and fourth models are a fuzzy design and a non-equivalent group, where assignment is based on non-random decision factors. For these two cases, $E[D_i|x_i]$ oscillates between
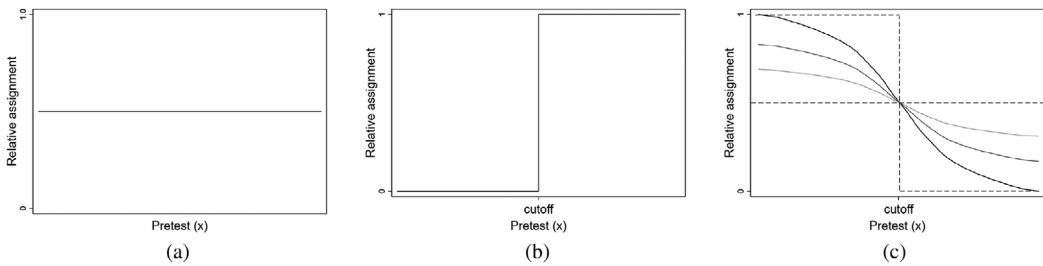
**FIGURE 3** Conditional probability of treatment; as in Trochim and Spiegelman (1980)

the horizontal line of the true experiment and the step function of the sharp design, as depicted in Figure 3c.

In most cases, the "relative assignment variable" successfully produces unbiased estimates. However, it is unclear whether results hold under more realistic or complex scenarios, for example when considering pretest-posttest nonlinear relationships or non-random assignment close to extreme values of the pretest distribution.

## 3.3 | Second formalization wave

At the turn of the millennium, Hahn et al. (1999) is one of the first studies that formally conducts a sharp RDD, allowing for variable treatment effects. The authors evaluate the effect of a federal anti-discrimination law on firm employment of minority workers, using demographic data from the National Longitudinal Survey of Youth. Specifically, the authors define the running variable as the number of employees of a given firm, and consider treatment as any firm with 15 employees or more, which was covered by the law. Within their findings they show that, for at least 2 years, the Equal Employment Opportunity Commission coverage had a positive effect on the percentage of minority workers employed by small firms.

We document the work of Angrist and Lavy (1999) as one of the first formal applications to fuzzy designs. The authors set their identification strategy around the Maimonides' Rule (named after the 12th-century rabbinic scholar Maimonides) to study the effects of class size on students' achievements in Israel. Formally, the rule is characterized as follows:

$$f_{sc} = e_s/[int((e_s - 1)/40) + 1] \tag{10}$$

where $f_{sc}$ denotes the number of students in classroom $c$ of school $s$, $e_s$ is the beginning-of-the-year enrollment, and the function $int(n)$ corresponds to the largest integer less than or equal to $n$. Hence, enrollment cohorts of 41–80 are split into two classes of average size 20.5–40, cohorts of 81–120 are split into three classes of average size 27–40, and so on. Figures 4a and 4b show the actual class size compared to the Maimonides' Rule for 2 years in elementary school (4th and 5th grade, respectively), as used by Angrist and Lavy. As shown, there was imperfect compliance to the rule, due to the many factors involved. For instance, schools with a high Percent Disadvantage Index receive funds that are sometimes earmarked for additional classrooms.

As such, Angrist and Lavy conduct an instrumental variable approach, common in the recent RDD setup (see Section 2.1), with the following first and second stages:

$$\text{First Stage: } n_{sc} = X'_s \pi_0 + f_{sc} \pi_1 + \epsilon_{sc} \tag{11}$$
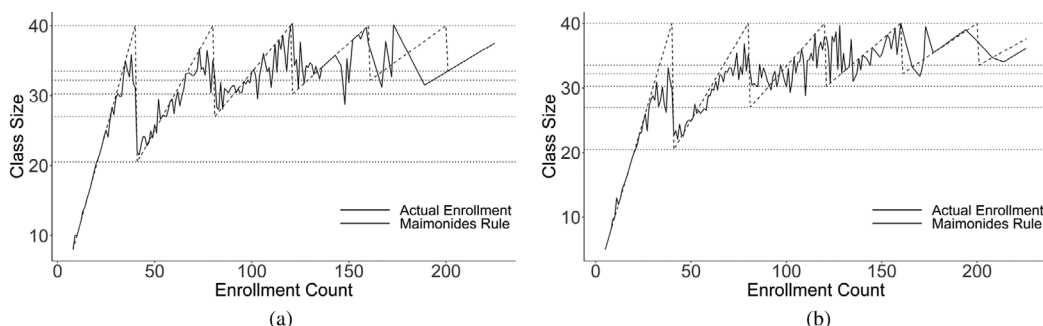
**FIGURE 4** Class size in 1991 by actual size and as predicted by Maimonides' Rule in (a) Fourth and (b) Fifth grade. Replication made using data from Angrist and Lavy (1999)

$$\text{Second Stage: } \bar{y}_{sc} = X'_s\beta + n_{sc}\alpha + \eta_s + \mu_c + u_{sc} \tag{12}$$

where $n_{sc}$ denotes the actual classroom size, $f_{sc}$ is the class size as dictated by Maimonides' Rule, and $X'_s$ is a vector of school-level covariates that include functions of enrollment, student's socioeconomic status, as well as variables identifying the ethnic character and religious affiliation. In turn, $\bar{y}_{sc}$ is the average class score, and $\eta_s$, $\mu_c$ capture within-school and within-class correlation in scores. As a result, Angrist and Lavy find a negative effect of class size on test scores for mathematics and reading scores. More specifically, they find that smaller classes (i.e., classes with 10 fewer students) achieve a 75% higher grade score.

We finally document the work of Leuven and Oosterbeek (2004) who conduct both sharp and fuzzy designs to evaluate the effects of tax deductions on training participation, and the effects of training participation on wages, respectively, for the case of the Netherlands. Essentially, the authors exploit a tax deduction, enacted in 1998, for firms that trained employees aged 40 years or more. This deduction translates into a cost discontinuity, with a significantly lower cost of training someone right at the age of 40 years than someone right below 40. Hence, to evaluate the effect of tax cuts on training participation, Leuven and Oosterbeek estimate the following model:

$$E(t_i) = \alpha + \beta d_i \tag{13}$$

where $t_i$ denotes the probability of receiving training and $\alpha$ is the probability of receiving training without a tax deduction. The coefficient of interest, $\beta$, is interpreted as the change in training probability due to the extra tax deduction in the vicinity of $\bar{a} = 40$ years of age, that is, $\lim_{\epsilon\downarrow 0} E(t \mid a = \bar{a} + \epsilon) - \lim_{\epsilon\uparrow 0} E(t \mid a = \bar{a} + \epsilon)$.

Alternatively, the authors evaluate the effects of training participation on wages. Given that the probability of training does not necessarily match the actual realization of being trained, they estimate the following fuzzy design:

$$E(w_i) = \omega + \gamma(t_i) \tag{14}$$

where $\omega$ denotes the wage (in logs) without training, and $t_i$ is instrumented with the probability of treatment as a function of age, $Pr(t_i) = f(a_i, 1[a_i \geq \bar{a}])$. The coefficient of interest, $\gamma$, is interpreted as the localized change in wages due to training. Results in Leuven and Oosterbeek (2004) show
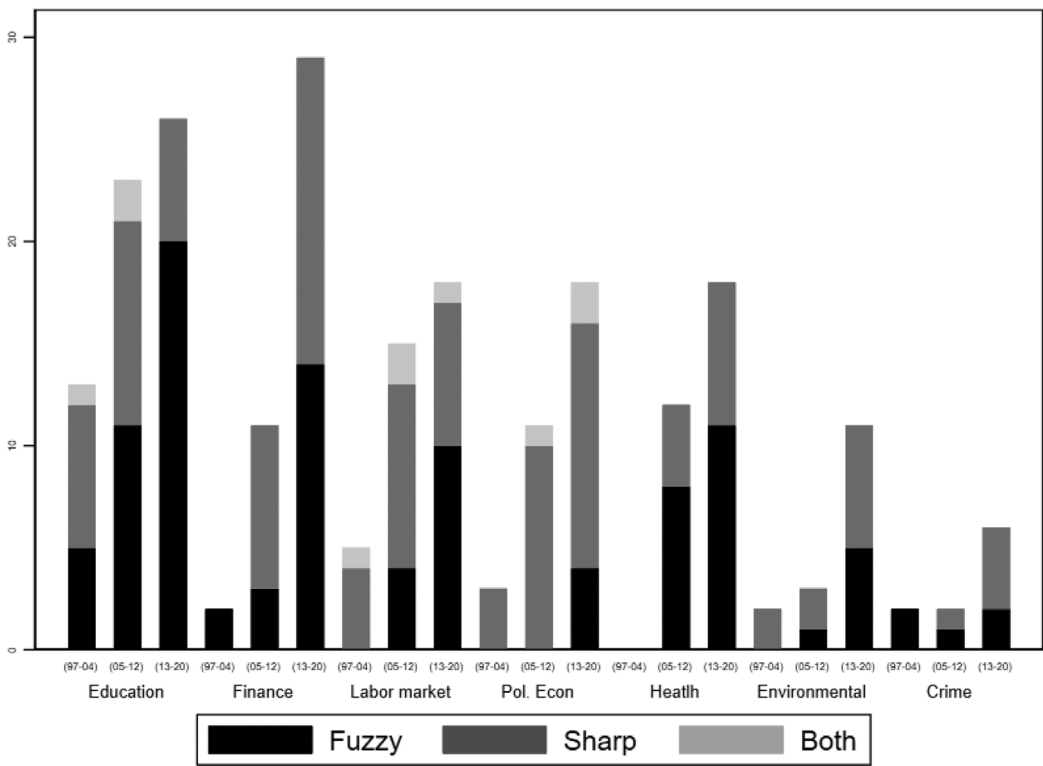
**FIGURE 5**    Empirical studies by economic field and type of design: 1997–2020

that employees just over 40 years of age have 15%–20% higher probability of training than those below 40 years. However, they find a null effect of participation on wages.

## 4 | RDDS BY ECONOMIC FIELD (*IN ALPHABETICAL ORDER*)

In this section we break the RDD literature down by economic field, highlighting the main outcomes, treatments, and running variables employed. We also report the historical distribution of fuzzy versus sharp designs and provide some intuition as to why some criteria produce imperfect compliance (cases in which treated and control observations were partially enacted). Most of our analysis in this section is based on literature from the new millennium, since we note that only nine empirical RDD studies were conducted before the year 2000 (mostly centered on issues pertaining to education and political economy).

As depicted in Figure 5, during 2005–2012, we observe over 76 studies, covering a wider range of fields, including: education (23), labor market (15), health (12), political economy (10), finance (11), environment (3), and crime (2). Notably, the number of sharp designs during this period nearly doubled the number of fuzzy designs. The last time period dates from 2013 until 2020 and shows an even larger increase in the number of empirical RDD applications, with close to 132 studies, most of which center on finance (29) and education (26).

Overall, we see some topics in economics gaining importance through time, like the cases of health, finance, crime, environment, and political economy. Nonetheless, education stands out as

**T A B L E 1**    Empirical studies by sharp or fuzzy design: 1997–2020

| Sorted alphabetically | Fuzzy | Sharp | Both | Total |
|---|---|---|---|---|
| Crime | 5 | 5 | 0 | 10 |
| Education | 36 | 23 | 3 | 62 |
| Environment | 6 | 10 | 0 | 16 |
| Finance | 19 | 23 | 0 | 42 |
| Heatlh | 19 | 11 | 0 | 30 |
| Labor market | 14 | 20 | 4 | 38 |
| Others | 3 | 3 | 0 | 6 |
| Pol. Econ. | 4 | 24 | 3 | 31 |
| Total | 106 | 119 | 10 | 235 |

**T A B L E 2**    Common variables in crime

| Relative frequency | Common outcomes | Common treatment and running variable | Fuzzy/Sharp |
|---|---|---|---|
| 18% | Inmates misconduct and Arrest rates | Prison security levels ↔ Classification score | Fuzzy |
| 18% | Post-release criminal behavior | Severity of punishments ↔ Age at arrest and adjudication score | Fuzzy/Sharp |
| 18% | Juvenile crime | Delayed school starting age ↔ Birthdate | Fuzzy/Sharp |

the uncontested champion through time, with a total number of 62 empirical studies out of a total of 235.

## 4.1 | Crime

In the area of crime, RDDs have been applied in research topics related to the evaluation of prisons' risk-based classification systems, the effects of incarceration, prison conditions, and the deterrence effect of criminal sanctions, among others. In this field we acknowledge the pioneer and seminal paper by Berk and Rauma (1983), who examine unemployment benefits to prisoners after their release. Also, in Table 2 we document popular treatments that range from prisons' security levels and the severity of punishments to educational levels. Outcome variables generally focus on the number of arrests, inmates' misconduct, post-release behavior, and juvenile crime.

We find that roughly 50% of the crime literature employs fuzzy designs (see Table 1), with date of birth and prisoners' classification scores as commonly used running variables. For instance, Berk and de Leeuw (1999) evaluate California's inmate classification system, where individual scores are used to allocate prisoners to different kinds of confinement facilities. For example, "*sex offenders are typically kept in higher-security facilities, because a successful escape, even if very unlikely, would be a public relations disaster*" (p. 1046). The *fuzziness* in this case study stems from administrative placements decisions (different from the score), based on practical exigencies, such as too few beds in some facilities or too many in others.

Also, Lee and McCrary (2005) evaluate the deterrence effect of criminal sanctions on criminal behavior by comparing outcomes around the 18 year-age-of-arrest threshold. In particular, the

authors exploit the fact that offenders are legally treated as adults the day they turn 18 and hence are subject to more severe punitive adult criminal courts. However, all states in the US have the option to transfer a juvenile offender to a criminal court to be treated as an adult (according to the authors, this is particularly the case in Florida). Thus, the study corrects treatment assignment with a fuzzy framework.

### 4.1.1 | Deterrence effects of incarceration

The crime literature generally establishes that increasing the severity of punishments deters criminal behavior by raising the "expected price" of committing a crime. This approach is supported by Becker (1968), who was one of the first authors to claim an inverse relationship between criminal behavior and the likelihood of punishment. Henceforth, several authors have been interested in providing empirical evidence indicating that court sanctions are effective in reducing criminal behavior.

A more recent example is Hjalmarsson (2009), who analyzes the impact of incarceration on the post-release criminal behavior of juveniles. The author exploits institutional features of the justice system, namely discontinuities in punishment that arise in the state of Washington's juvenile sentencing guidelines. More specifically, these guidelines consist of a sentencing grid that establishes punishment based on the severity of the offense and the criminal history score. In this setting, an individual is sentenced to a state facility for a minimum of 15 weeks if he/she falls above a cutoff, otherwise, the individual receives a minor sanction (i.e., fine or probation). Results in Hjalmarsson indicate that incarcerated individuals are less prone to be re-convicted of a crime.

Additionally, Lee and McCrary (2005) investigate the deterrence effect of jail-time on criminal behavior. The underlying endogeneity problem is that, by design, severe criminal sanctions receive lengthier times of incarceration. To overcome this problem, the authors exploit the fact that when an individual is charged with a crime before his 18th birthday, his case is handled by juvenile courts. In contrast, if the offense is committed during or after his 18th birthday, it is handled by the adult criminal court, affronting a more punitive criminal sanction. Surprisingly, Lee and McCrary find that individuals are greatly unresponsive to sharp changes in penalties. This result is consistent with extremely impatient or myopic individuals, who essentially are unwilling to wait a short amount of time to see their crime sentence significantly reduced.

### 4.1.2 | Prisoners' placement

With a growing number of convict populations and ever tighter budget constraints, prison systems have been looking for measures to improve their efficiency. And, a key criterion for an effective placement is defined by the capability of controlling inmates' misconduct. We find several authors that provide an evaluation of how well the system currently allocates inmates to incarceration facilities.

For instance, Berk and de Leeuw (1999) provide an evaluation of the inmate classification system used by the State of California. Namely, the Department of Corrections estimates a classification score that captures the likelihood of potential misconduct. This likelihood includes a linear combination of controls, such as an inmate's age, marital status, work history, and above all, the length of the sentence, which accounts for almost 70% of the variance in the overall classification score. After the score is computed, placement is undertaken. Berk and de Leeuw conclude that

the existing scores successfully sort inmates by different levels of risk, and thus, the assignment based on sorted security-level facilities reduces the probability of misconduct.

Similarly, Chen and Shapiro (2003) evaluate the effect of prison conditions on recidivism rates, exploiting discontinuities in the assignment of prisoners to security levels based on a score that reflects the need for supervision. The authors use a sample of 1205 inmates released from federal prisons in the year 1987, with data on physical and social conditions of confinement as well as post-release criminal activity. The authors find that harsher prison conditions significantly increase post-release crime.

### 4.1.3 | Crime and other economic fields

We finally document some studies that center on the causal relationship between crime and other economic fields. Berk and Rauma (1983) for instance, evaluate ex-offenders' eligibility for unemployment benefits on reincarceration rates. That is, inmates in the state of California are eligible for unemployment benefits if they enroll in vocational training programs or work in prison jobs. The authors compare individuals with and without unemployment benefits after their release and find that benefits effectively reduce recidivism.

In turn, Depew and Eren (2016) explore the link between school entry age (as previously reported in Section 4.1) and crime. Specifically, the authors compare incidences of juvenile crime in the state of Louisiana, by neighboring in on the interaction between birth dates and official school entry cutoff dates, and find that late school entry reduces the incidence of juvenile crime for young black females, particularly in high crime areas. Somewhat related, Cook and Kang (2016) investigate whether there is a causal link between student dropouts and crime rates. The authors first demonstrate that children born just after the cutoff date for enrolling in public kindergarten are more likely to drop out of high school. The authors then present suggestive evidence that dropout precedes criminal involvement.

## 4.2 | Education

While studies in the field of education cover an ample range of issues, most however relate to student performance. As Table 3 shows, most studies (66%) center on either academic achievement or educational attainment while others (19%) center on student enrollment or access to higher education. Common treatment variables for the former include summer school attendance, class size, a delayed school starting age, and monetary incentives to teachers. On the other hand, treatments for enrollment and access to higher education include attendance and financial aid.

We observe that almost 58% of the entire education literature employs fuzzy designs (see Table 1). This is particularly the case for studies that choose *test scores* as a running variable. In Jacob and Lefgren (2004b) for instance, remedial programs were assigned to students scoring below a passing grade. However, the authors note that some failing students received waivers from summer school and some passing students were retained due to poor attendance. In Jacob and Lefgren (2004a), the same authors evaluate the effects of teacher training in schools that were placed on probation under the Chicago Public System. Among the several estimation challenges, the authors note that (i) student mobility (between probation and non-probation schools), (ii) schools that were later placed on probation, and (iii) schools that were initially on probation but managed to be removed, constitute key reasons to consider a fuzzy discontinuity framework.

**TABLE 3** Common variables in education

| Relative frequency | Common outcomes | Common treatment and running variable | Fuzzy/Sharp |
|---|---|---|---|
| 66% | - Academic achievement<br>- Educational attainment | Summer school attendance ↔ Test scores | Fuzzy |
| | | Class Size ↔ Student enrollment | Fuzzy |
| | | Delayed school starting age ↔ Birthdate | Fuzzy |
| | | Attending better schools ↔ Admission exam score | Fuzzy |
| | | Monetary incentives to teachers ↔ Program eligibility criteria | Sharp |
| 19% | - Enrollment<br>- Accessibility to higher education and school | Attending better schools ↔ Admission exam score | Fuzzy |
| | | Financial aid ↔ Program eligibility criteria | Sharp |

Similarly, Van der Klaauw uses a college eligibility criterion (i.e., a continuous variable based on academic ability) to evaluate the effects of financial aid on student enrollment. The author nonetheless provides reasons for why the criteria do not perfectly predict treatment. For one, college enrollment is influenced by a number of different factors, many of which are unobserved by college administrators. Also, observed characteristics such as awards, recommendation letters, and extracurricular activities are often hard to measure. Moreover, a college might offer additional aid to a weaker student if the probability of getting a better student is low.

Finally, we document the work of Samarakoon and Parinduri (2015) who evaluate the effects of education on women empowerment by exploiting a government policy requiring a longer school year in Indonesia in 1978. Essentially, students born in 1971 or before would experience a 6-month more extended school year. However, the possibility of student drop-outs or a potential lag-delay in school entry (e.g., a student born in 1971 that entered with the 1972 cohort) would introduce imperfect compliance to the birth-date design. The authors correct for these cases with a fuzzy framework.

### 4.2.1 | School quality

Angrist and Pischke (2009) argue that class size is one of the most expensive inputs in education, as it involves additional teachers and classrooms. As such, the Tennessee STAR experiment in 1986 was an ambitious 4-year randomized controlled trial (with an approximate cost of USD 12 million) that showed the benefits of having smaller classes for the first time. In the RDD literature, the seminal paper by Angrist and Lavy (1999), covered in Section 3, addresses the same question with similar findings. Other RDD studies that have followed this particular evaluation include the works of Hoxby (2000), Asadullah (2005), Urquiola (2006), and Urquiola and Verhoogen (2009).

Another influential paper during this time is Black (1999), which evaluates the quality of schools on students' outcomes. Given the inherent difficulty in calculating parents' willingness to pay on better schools, Black instead looks at the real estate market within school districts. Namely,

she compares differences in housing prices on opposite sides of attendance district boundaries and finds that parents are willing to pay 2.5% more for a 5% increase in test scores. Similar findings on the effects of "*magnet schools*", and their potentially higher peer quality, include the works of Pop-Eleches and Urquiola (2013), Park et al. (2015), Dee and Lan (2015), Anderson et al. (2016), and Wu et al. (2019).

The effect of better infrastructure and resource allocation in schools is also an issue that has gained importance over the last two decades. Chay et al. (2005), for example, show that the impact of government programs that are earmarked for infrastructure is largely overstated. The authors use Chile's 900 Schools Program as a case study, where participation was strongly determined by whether a school's mean score fell below a particular value relative to its region. Treated schools received infrastructure improvements, instructional materials, teacher training, and tutoring for low-achieving students. Interestingly, the authors find a significant effect during periods in which the program was not yet operational and also a sharp decline in average test scores in the year used to assign program participation. Hence, while Chay et al. conclude that the program did have an impact on scores, they show that many favorable reviews were simply an artifact of mean reversion. For additional literature on the effects of infrastructure in schools, we refer readers to Cellini et al. (2010), and Hong and Zimmer (2016). Also, studies that center on the benefits of increased educational funding include those of Guryan (2001), Leuven et al. (2007), and Van der Klaauw (2008). In general, these studies cover government programs that allocate resources based on state aid formulas, poverty indices, and percentage of disadvantaged minority pupils.

## 4.2.2 | Educational leveling

A substantial part of the literature dealing with educational leveling focuses on the effects of remedial programs. Such is the case of Aiken et al. (1998), who evaluate a university-level (freshman) remedial writing program based on the ACT English placement test score. The authors measure the effects on two writing outcomes: a written essay (indirect assessment) and the Test of Standard Written English -TSWE (direct assessment). Among their findings, the authors show a positive effect of remediation on TSWE scores when using the ACT as a running variable, but find a non-significant effect when using the Suite of Assessments -SAT scores as a running variable.

Remedial programs based on summer schools and grade retention are queenly studied in Jacob and Lefgren (2004b). In particular, the authors evaluate an accountability policy of Chicago Public Schools enacted in 1996, which tied remedial programs to standardized tests scores. The authors find a positive and lasting effect (of up to 2 years) of summer schools in reading and mathematics, but find mixed effects of grade retention (if anything appearing to increase performance of third graders in the short run).

The same authors study the effects of in-service training in schools. In Jacob and Lefgren (2004a) they exploit an educational reform enacted by the Chicago Public System in 1996, where 71 of its 489 elementary schools were placed on academic probation. Namely, schools in which less than 15% of students scored above national norms (in reading) were subject to probation and thus eligible to receive funding for staff development, technical assistance, and enhanced monitoring. Surprisingly, the authors do not find a statistically significant effect of in-service training on either reading or math achievements, and postulate that modest investments are not enough to improve achievements in poverty schools.

### 4.2.3 | Enrollment

Apart from programs designed to improve the quality of schools and remedial programs intended to level students' performance, there is also a strand of the literature that focuses on students' enrollment. That is, programs that promote access to students that, mostly because of financial needs, would not have had the possibility to enroll. A pioneer work on the matter is Van der Klaauw (2002), who evaluates the effects of a rule-based financial aid policy based on a measure of academic ability used by an anonymous college in the United States East Coast during 1989–1993. The author finds that the enrollment elasticity with respect to college grants was 0.86 and 0.13 for students who had applied for federal aid and for students who were ineligible for federal aid, respectively. Van der Klaauw (2002) largely paved the way for similar research, including the works of: Kane (2003), Canton and Blom (2004), Goodman (2008), Filmer and Schady (2009), and Barrera et al. (2013).

An interesting exercise is proposed in Meyersson (2014), which studies the effects of Islamic political representation on female secular high school completion. The author centered his investigation in Turkey around the 1994 political change when the pro-Islamic Refah Party gained an unprecedented representation (the second-largest party in terms of votes). Meyersson finds positive effects (close to a relative increase of 20%) on female enrollment and high school completion. In the same context, Önder and Shamsuddin (2019) exploit the multi-cutoff heterogeneity in the 1994 voting shares and find similar results. Nonetheless, Önder and Shamsuddin argue that variations in their results are possibly associated with a policy change in 1999 that restricted the number of religious high school students entering university.

A recent work by Goodman et al. (2019) evaluates whether online courses increase access to formal education. Given the novelty of online schooling, it is at the forefront of studying its effect on enrollment. In particular, Goodman et al. exploit the way in which the Georgia Institute of Technology's Online M.S. in Computer Science admitted students, that is, applicants for the 2014 summer course were required a Grade Point Average - GPA of 3.26 or above. In short, the authors find that all (100%) of the marginal admits to the program later enter into formal higher education.

### 4.2.4 | Years of schooling

We finally turn to the literature that centers on the duration of schooling. In most cases, authors exploit regulations where individuals born right before a school entry cutoff date carry an additional year of education compared to those born right after. In those cases, treatment within similar age cohorts is almost "*as good as randomly assigned*". Several studies evaluate the impact of this additional year on various measures of academic performance, such as the Armed Forces Qualifying Test Scores of minorities (Cascio & Lewis, 2006), test scores at kindergarten or high school entry (Datar, 2006, Zhang et al., 2017), and school enrollment (McEwan & Shapiro, 2008).

Other studies evaluate the effect on longer-term variables, such as health care services (Dang, 2018). Also, Samarakoon and Parinduri (2015) explore the effects on women empowerment, by focusing on a regulatory reform in the schooling system of Indonesia. Specifically, in order to synchronize the academic year with the government budget year, the Indonesian Minister of Education and Culture required schools to lengthen the 1978 academic year up until June 1979. Hence, children who began their school year in January of 1978 completed their coursework in June 1979 (after 18 months), rather than in December of 1978.

**TABLE 4** Common variables in environment

| Relative frequency | Common outcomes | Common treatment and running variable | Fuzzy/Sharp |
|---|---|---|---|
| 44% | - Mortality<br>- Life expectancy<br>- Production<br>- Housing prices | Regulatory status (Air quality improvements) ↔ Pollution levels | Fuzzy |
| 25% | Air pollution | Driving restrictions ↔ Date | Sharp |

In principle, this meant that a student born in 1971 or before would experience a 6-month longer school year. As a result, Samarakoon and Parinduri find a reduction in the number of live births, an increase in contraceptive use, and promotion in reproductive health practices.

## 4.3 | Environment

In the field of environment, most of the research has centered on two affairs (see Table 4). One is the impact of policies that induce air quality improvements on outcomes such as mortality, life expectancy, and housing prices. The other deals with the effectiveness of driving restriction policies on air pollution levels. We document that nearly 38% of the environmental studies in our sample employ fuzzy designs. This is for example the case of Greenstone and Gallagher (2008), who estimate local welfare impacts of the Superfund-sponsored cleanups of hazardous waste sites. Particularly, the authors implement a quasi-experiment based on the selection rule that the Environmental Protection Agency (EPA) used to develop the National Priorities List (NPL). It follows that EPA had only enough money to conduct 400 cleanups, which were chosen based on the highest scores of the Hazardous Ranking System (HRS). As such, the authors compare housing prices in areas surrounding the selected waste sites to areas surrounding other waste sites that narrowly missed qualifying for a cleanup. The *fuzziness* in this study stems from the fact that some areas below the HRS threshold received cleanups. The authors do not find any statistically significant change in residential property values, rental rates, and housing supply.

Another example is Zhang et al. (2018), who evaluate the effectiveness of central supervision on industrial firms' chemical oxygen demand (COD) emissions. The authors rely on the National Specially Monitored Firms (NSMF) program that imposes direct supervision on polluting firms above a specific COD and NH3-N emission threshold. However, roughly 20% of highly polluting firms received waivers while 10% of firms below the cutoff were monitored. Under this fuzzy framework, the authors find that central supervision significantly reduced industrial COD emissions by at least 27%.

### 4.3.1 | Driving restrictions and tolls

Motor vehicles are notorious for their negative externalities, including traffic and air pollution. As such, driving restriction policies and tolls have been adopted in numerous cities worldwide. However, addressing the cost-effectiveness of these policies is challenging mostly due to concurrent policies that affect air quality. In this literature, Davis (2008) evaluates the effect of the *Hoy No Circula* (no transit today) program in Mexico City, which restricted drivers from using their

vehicles 1 day per week. Using hourly measures of the five major air pollutants, the author compares air quality before and after the restriction to identify causal links. Surprisingly, he finds no evidence that the program improved air quality. Instead, there is evidence that the restriction led to an increase in the number of vehicles in circulation. Also, Davis finds evidence of inter-temporal driving substitution towards evenings and weekends.

Similarly, Viard and Fu (2015) evaluate pollution effects from two driving restrictions in Beijing. Namely, an odd-even license plate system began in July 2008 that prevented driving every-other-day. Later, in October 2008, the government re-instated a less restrictive system preventing driving 1 day per week. The authors find that the every-other-day restriction reduced particulate matter by 18% and the 1-day-per-week restriction by 21%. In contrast to Davis (2008), the authors find little evidence of intertemporal substitution of driving. A very similar case is studied by Ye (2017) who uses Lanzhou as case study, one of the most polluted cities in northwestern China. In this case, driving restrictions were strengthened in 2013, by moving from a 1-day-per-week to an every-other-day restriction. This policy change was directed at having less traffic during a series of special events, such as the Lanzhou Investment and Trade Fair and the Lanzhou International Marathon Games. Ye find that the restriction upgrade (from 1-day to every-other-day) was overall ineffective in improving air quality, most likely due to the hiring of alternative vehicles or higher car sales.

Fu and Gu (2017) are among the few to study the use of tolls as a policy to alleviate congestion and to curb vehicle carbon emissions. The authors exploit the fact that tolls were waived nationwide in China for passenger vehicles during the 8-day National Day holiday in 2012. Their findings suggest that eliminating tolls increases pollution by 20% and decreases visibility by 1 km.

### 4.3.2 | Other environmental policies

Several authors exploit regulatory policies to study a wide range of environmental issues. Such is the case of Chay and Greenstone (2003) who examine the 1970 Clean Air Act Amendments (CAAA) in the United States and its effects on infant health. Specifically, the legislation imposed more stringent regulations on counties where industrial polluters exceeded a threshold of concentrated suspended particles (TSPs). By comparing counties close to the threshold, they find that a 1% decrease in TSPs yields a 0.5% decline in infant mortality rate. Chay et al. (2003) also use the CAAA to examine effects on adult health across nonattainment and attainment counties. They find that in the first year post-implementation of the Amendments, TSP concentrations declined significantly in nonattainment counties. However, they find scant associative evidence between the nonattainment status and changes in mortality rates for both adults and the elderly. Later, Chay and Greenstone (2005) exploit the CAAA to determine the impact on the real estate market and report a housing price elasticity of TSP concentrations in the range of -0.20 to -0.35.

Another example of TSPs is presented in Chen et al. (2013) who evaluate their effects on life expectancy. The authors exploit the Chinese Huai River policy, which provided free winter heating via the provision of coal for boilers in cities north of the Huai River but denied heat to the south. In this sense, the authors exploit the discontinuous increase in the availability of free indoor heating as an individual arbitrarily crossed the Huai River line. The authors demonstrate that free heating comes at a high price: a loss of 5.5 years in life expectancy.

Finally, we document an environmental policy dealing with oil production regulations. In the study by Balthrop and Schnier (2016), the authors compare two neighboring US states (Texas and Oklahoma) based on regulatory differences. Namely, while Oklahoma places greater emphasis on securing property-rights through unitization, Texas has stricter oil well-spacing guidelines

**TABLE 5**  Common variables in finance

| Relative frequency | Common outcomes | Common treatment and running variable | Fuzzy/Sharp |
|---|---|---|---|
| 11% | - Corporate investment<br>- Financial policies<br>- Employment<br>- Innovation | Loan covenant violation ↔ Current ratio, net worth and tangible net worth | Sharp |
| 7% | - Output<br>- Labor supply<br>- Schooling<br>- Household expenditure and assets | Group-based micro credit programs ↔ Land tenure | Fuzzy |
| 5% | - Approved mortgage loans<br>- Housing prices | Judicial foreclosure requirement ↔ Distance from state boundary | Fuzzy |
| 3% | - Investment<br>- Turnover<br>- Employment<br>- Productivity | Receive subsidy (firms)↔ Score | Sharp |

and severance taxes. Hence, the authors test whether Oklahoma is more successful in terms of cumulative physical recovery over the lifetime of an average well. To do so, they use geographical coordinates with respect to the border and along the Anadarko Basin. Among their results, the authors find that Oklahoma produces 3361 more barrels of oil over the lifetime of an average well.

## 4.4 | Finance

In the area of finance we document that most of the research has centered on the effects of credit default on variables such as corporate investment, financial policies, employment, and innovation (see Table 5). Several authors have studied the impact of microcredit programs for the poor on outcomes such as output, labor supply, schooling, and reproductive behavior in low-income countries. Also, some authors exploit differences in foreclosure laws across state boundaries on mortgage loans and housing prices. Roughly 45% of the work in finance employs a fuzzy framework. For instance, Chang et al. (2014) investigate the price effects of stock market indexing. In particular, the Russell 1000 and 2000 stock indices incorporate the first 1000 (1–1000) and the next 2000 (1001-3000) largest firms ranked by market capitalization, respectively. The authors exploit the fact that marginal changes in capitalization of firms close to the 1000th rank, move them *as good as randomly* between indices. That is, small changes in market capitalization lead to large and discontinuous changes in demand. Nonetheless, the authors employ a fuzzy RDD since their measure of market capitalization does not perfectly match the one used by Russell Inc. Put differently, measurement error in the running variable produces an imperfect index-membership sorting. Among their findings, Chang et al. indicate that an index demotion (addition to the Russell 2000) causes a price increase while an index promotion causes a price reduction. The reason is that indices are value-weighted, where more money tracks the largest stocks within each index.

Hence, *it is better to be a big fish in a small Russell 2000 pond than a small fish in a big Russell 1000 pond*.

Another example is Garmaise and Natividad (2017) who study, for the Peruvian case, the impact of unfavorable credit events on consumers' long-term borrowing. More specifically, consumers with a mix of different currency denominated debt can ultimately face a credit rating downgrade due to an exogenous shock (in this case, bad luck due to exchange rate movements). Additionally, by Peruvian banking regulations, "*a poor risk rating given by any bank with a share of 20% or more of a given borrower's total lending should be reflected in the ratings of all other lenders*". Hence, the authors compare borrowers with banking relationships whose balances barely exceed 20% with borrowers with banking relationships whose balances just miss the threshold. As a result, Garmaise and Natividad find that the credit downgrade accounts for a 25%–65% decline in borrowing, at horizons of up to 3 years.

Also, we document Garmaise (2013) who evaluates the impact of a mortgage program in the United States intended to increase borrowers' flexibility in the timing of their repayments. That is, banks can offer standard loans to any borrower (with potential annual payment increases) or flexible loans (with fixed payments) to applicants with a credit score above a given threshold. However, some formally eligible borrowers received standard loans, and some ineligible borrowers received flexible loans. Results under a fuzzy framework show that the program increased the bank's volume of originations by over 35%. However, it attracted borrowers who were four times as likely to experience credit delinquency (i.e., non-performing loans).

Finally, we highlight a novel wave of studies related to applications of RDDs on Fintech. For instance, Cumming et al. (2021) present the first evidence that marketplace credit causes a significant increase in the level of entrepreneurship. Specifically, using a fuzzy RDD that exploits exogenous variation in borrowers' access to marketplace loans across US state borders, the authors show that a 10% increase in marketplace lending leads to a 0.44% increase in business establishments per capita. Also, we find some research related to the effects of access to Fintech loans. This is the case of Bharadwaj et al. (2019), who examine the take-up and impact of one of the most popular digital loan products globally, M-Shwari in Kenya. The authors conclude that digital loans improve financial access. Similarly, Chen et al. (2020) analyze the effect of Fintech credit access on micro and small firm volatility.

### 4.4.1 │ Loan covenant violations

Several authors have been interested in the impact of debt covenant violations, otherwise known as "technical defaults" (i.e., the violation of any covenant other than the payment of interest or principal of the loan), on corporate investment. Chava and Roberts (2008) for example, are among the first to identify the mechanism through which creditors gain control rights of the borrowing firms when breaching a covenant. Under these situations, creditors can either threaten to accelerate payment or ultimately intervene in management decisions to ensure a safer return on their investment. Chava and Roberts find that capital investment declines by 13% relative to investment levels prior to covenant violations and show that effects are magnified when agency and information problems are more severe.

Mariano and Giné (2015) contribute to this analysis by considering firm's growth opportunities at the time of the violation. On the one hand, growth opportunities are often a source of conflict between shareholders and creditors, since investment decisions may lead to situations in which creditors are unlikely to get their money back (Jensen & Meckling, 1976). On the other

hand, growth opportunities constitute an important source of future income flows, which in turn can be necessary to meet outstanding debt obligations. Following a covenant violation, the authors find that investment declines (increases) when the firm has few (many) growth opportunities. Also, the firm's performance improves but with dividend cuts, and the frequency of CEO turnover increases.

Similarly, Roberts and Sufi (2009) demonstrate that conflicts between firms and creditors have a large impact on corporate debt policy. Falato and Liang (2016) provide evidence of these adverse effects on employment, that is, sizable job cuts following a loan covenant violation. Finally, Gu et al. (2017) evaluate the effects on firms' innovation, showing that only the quantity (not quality) of innovation is negatively affected.

### 4.4.2 | Microcredit programs

Governmental and non-governmental agencies in low-income countries have introduced microcredit programs targeted to the poor in order to promote self-employment, income growth, and welfare. Particularly, rural finance has been an important tool for rural development and poverty reduction. However, measuring the impact of microcredit programs is challenging due to identification problems that are mostly caused by selection bias from governments' targeted interventions.

Pitt and Khandker (1998) are probably the first to study the heterogeneous effects of three group-based lending programs in Bangladesh, namely: *Grameen Bank*, *BRDB*, and *RD-12*. The authors exploit a financial regulation that restricts households that own more than 0.5 acres of land from enrolling in any of the three credit programs. Thus, the authors compare households around the land ownership eligibility threshold, and find that the program has a larger effect on women. For example, consumption expenditure for women increases by 18 *taka* compared to 11 *taka* for men, for every 100 additional *taka* borrowed.

Later, using data from villages in Bangladesh, Pitt et al. (1999) estimate the effects of group-based lending programs on reproductive behavior employing a very similar identification strategy. The central hypothesis is that lending programs affect fertility and human capital investments in children, by altering economic incentives through credit provision. The authors do not find evidence that women's participation in lending programs increases the use of contraceptives, nor in reducing fertility. In contrast, they find that men's participation reduces fertility and slightly increases contraceptive use.

Finally, Aung et al. (2019) provide insights into a large-scale subsidized credit program in Myanmar (formerly Burma), one of the poorest and isolated countries in Asia. Once again, the identification strategy relies on a credit provision rule, in this case, based on rice landholding size. Specifically, financial regulations dictate a maximum of 10 acres for farmers to borrow a fixed amount per acre of rice landholding. While the authors find little evidence that rice yield or output increases, they do find that the program increases household income, possibly due to a spillover effect on other farm activities.

### 4.4.3 | Housing market regulations

In the context of mortgage loans, economists and policymakers have focused significant attention on effective regulations to curb unaffordable mortgage debt. For instance, Pence (2006)

empirically assesses the effects of judicial foreclosure requirements on the size of approved mortgage loans. More precisely, she compares approved mortgages in census tracts that lie in close proximity but in different states. This approach is very similar to Black (1999) or Holmes (1998), and in this way, the author can control for unobserved variables that vary with location while identifying the effect of different judicial foreclosure requirements from the change at the state boundary. The author finds that loan sizes are 3%–7% smaller in states that require judicial foreclosure processes.

Employing a similar approach, Mian et al. (2015) investigate the effect of foreclosures on house prices and real activity during the last Great Recession in 2008–2009. Once again, the authors isolate causal effects by using differences in state laws in the foreclosure process. That is, a zip code-level analysis exhibits a discontinuous jump in foreclosure as one moves from a judicial to a non-judicial state. Additionally, some states require that a foreclosed sale must take place through the judicial courts. The authors find a large negative effect on housing prices due to the foreclosure-induced increase in housing supply.

Finally, we document the work of Kumar (2018) who exploits a 1997 constitutional amendment in the state of Texas, limiting home equity borrowing to 80% of the total home value. Hence, the author estimates the implications on overall leverage and mortgage defaults around the Texan border. Namely, he compares mortgage default probabilities or hazard rates of similar individuals located in neighboring counties but on either side of the state border. And, in contrast to the standard one-dimensional RDD setup, the author includes latitude and longitude spaced discontinuity coordinates, following the works of Dell (2010) and Keele and Titiunik. Among his findings, Kumar shows that home equity restrictions in Texas lowered the likelihood of default for residential mortgages between 2007 and 2011.

### 4.4.4 | Procurement auctions

Only about a handful of studies have viewed procurement auctions through a regression discontinuity lens, where bidders that bid close to the cutoff price differ discontinuously in winning or losing the auction. Among the few, Kawai and Nakabayashi (2014) analyze, for the case of Japan, collusion strategies among construction firms. Specifically, they first focus on failed auctions, where the bids of the lowest and second-lowest firms are close to each other. Under competition, these two firms are likely to have the lowest bid in a second auction, while collusion would still predict persistence. The authors identify close to 1000 firms with a conduct consistent with collusive behavior. Similarly, Kong (2021) studies the effects of sequential auctions with synergy, by comparing "*the second-auction behavior of a first-auction winner and first-auction loser who bid the same amount in the first auction*". Essentially, she evaluates whether correlation in bids is explained by affiliation or synergy. Among her findings, Kong indicates that affiliation is primarily responsible for the observed allocation patterns.

In turn, Bonaldi and Villamizar-Villegas (2018) study inventory effects of procurement auctions within the context of a private versus common-value test, as defined in auction theory. The authors compare bidders within the vicinity of the auctions' cutoff-price (barely winners and losers), and employ a fuzzy design to estimate the slope of the marginal valuation as a function of asset holdings. They find significant albeit small effects of auctions on the subsequent buying prices of the asset, which in their case amounts to the price of foreign currency.

We finally document Kuersteiner et al. who are the first to evaluate the effectiveness of foreign exchange intervention (FXI) using an RDD approach. The authors exploit a rule-based

**TABLE 6** Common variables in health

| Relative frequency | Common outcomes | Common treatment and running variable | Fuzzy/Sharp |
|---|---|---|---|
| 19% | - Health care utilization<br>- Out-of-pocket expenditure<br>- Mortality | Health insurance coverage ↔ Age (Index score) | Fuzzy/Sharp |
| 16% | - Subjective and objective health measures<br>- Health behavior<br>- Health care utilization | Retirement ↔ Age | Fuzzy |
| 13% | Infant mortality | An additional night in the hospital ↔ Time of birth | Fuzzy |
| | | Improved early life health care (VLBW classification)↔ Birth weight | Fuzzy |
| 6% | Health care utilization | Cost sharing reduction ↔ Age | Fuzzy |

intervention mechanism in Colombia during 2002–2012, where the central bank issued FX auctions if the exchange rate, relative to its monthly average, exceeded a given threshold. The authors thus compare episodes where procurement auctions took place with episodes where they barely missed the cutoff-rule. A methodological contribution of the paper is to extend RDDs from a cross-sectional to a time-series environment, and to show how these techniques can be used to identify local non-linear impulse response functions. The authors find a significant effect of sterilized interventions (i.e., a 1% depreciation in response to a 100 million US dollar purchase) that lasts for approximately 1 month. The authors also find that effects are amplified by the enactment of capital controls.

We note that Kuersteiner et al. (2018) later paved the way for other work on foreign exchange intervention using RDDs, including that of Perez-Reyna and Villamizar-Villegas (2019) and Vargas-Herrera and Villamizar-Villegas (2019).

## 4.5 | Health

In Section 2, we portrayed some of the close similarities between RDDs and Randomized Controlled Trials (RCTs). This close relationship has led authors like Lee and Lemieux (2010) to state that RDDs are "*a much closer cousin of randomized experiments than other competing methods*" (p. 289).

In the field of medicine, epidemiology, and health, RDDs are particularly useful since they can provide causal evidence on treatments, at low cost, especially in cases where there are little or no experimental trials. Yet, the RDD literature in these fields seems to be largely underutilized (Moscoe et al., 2015). In fact, we document that a significant portion of research has focused on the effects of health insurance coverage, commonly exploiting program eligibility criteria in both developing and developed economies. As shown in Table 6, other common treatments include the age of retirement, quality of health care, and patient cost-sharing reduction. Running variables generally center on patients' demographic characteristics, such as age and weight. In our sample,

nearly 63% of the health-related literature employs fuzzy designs, often benefiting from institutional setups (see Table 1). In most cases, studies use the fact that eligibility program status, retirement, minimum drinking age, patient cost-sharing reduction, etc., are triggered whenever individuals reach a certain age. The *fuzziness* however, is brought about by the decision of patients to take part. That is, age-triggering policies offer the right to treatment, but do not enforce treatment.

For instance, Zhang et al. (2018) evaluate the effect of retirement on health care utilization in urban China. The choice of retirement is addressed using a fuzzy framework, where the probability of retirement (optional) still maintains a discontinuous jump as a function of age. The authors find that retirement in China increases health care utilization.

Several medical inputs vary discontinuously across different measures of health risk. For example, birth weight has been commonly used as a running variable for the provision of better medical treatment. In this setting, Almond et al. (2010) compare health-related outcomes of newborns on either side of the Very Low Birth Weight (VLBW) threshold of 1500 g. Notwithstanding, birth weight is not the only factor used by doctors to provide special medical attention. Other variables that require immediate treatment to newborns include: respiratory rate, APGAR (Appearance, Pulse, Grimace, Activity, and Respiration) score, and head circumference. The cutoff of interest is hence fuzzy rather than sharp.

Clark and Royer (2013) constitute another example of a fuzzy design by shedding light on the rather elusive question as to whether education affects health. The authors use regulatory reforms in Britain's compulsory schooling laws, passed in 1947 and 1972, which lengthened the minimum age at which children were allowed to leave school. Basically, instead of students leaving school at the earliest opportunity, they had to turn 14 or 16 years of age, depending on the reform. However, educational attainment is not perfectly predicted by an individuals' birth cohort, due in part to the possibility of student drop-outs or a potential lag-delay in school entry (a similar case is documented in Samarakoon & Parinduri, 2015, in Section 4.1). As a result, while the authors confirm increases in completed years of education and earnings, they find little or no evidence on health effects.

Finally, we document the work of Dykstra et al. (2019) which exploits countries' eligibility criteria for Gavi: a global consortium with the goal of creating equal access to new and underused vaccines for children living in poor countries. Initially, the program excluded all countries with a per capita Gross National Income above US $1000. However, the authors recognize that once a country qualifies for aid, it maintains its eligibility, regardless of exceeding the per capita income cutoff. Additional features of the program that merited a fuzzy design included coverage thresholds for specific funding windows and the fact that many Gavi-eligible countries did not receive aid in a given year. Among their findings, the authors report that aid for cheap vaccines such as DPT and Hepatitis B were, if anything, inframarginal.

### 4.5.1 | Early-life health care

Several authors exploit rules that provide specialized medical care to children who are born underweight. One example is Almond et al. (2010) who argue that medical expenses in the United States are high and increasing. Thus, an estimation of whether the benefits of additional medical expenditures exceed their costs is remarkably relevant. Using data on the census of U.S. births, Almond et al. estimate marginal returns to medical care by comparing health outcomes and medical treatment provision for newborns at either side of the Very Low Birth Weight (VLBW) threshold of 1500 g. Patients above the threshold incur additional medical costs and the authors estimate the

associated benefits by examining average differences in health outcomes. The authors find that newborns weighing just below 1500 g have substantially lower mortality rates compared to those just above, despite a general decline in health associated with lower weight.

Barreca et al. (2011) reevaluate the effects of the VLBW classification on infant mortality provided by Almond et al. (2010). Namely, the Almond et al. (2010) study revealed extensive heaping (i.e., *bunching*) at 1 ounce and 100 grams multiples, in part related to technological constraints in measurement precision and tendencies to round numbers for convenience. Motivated by this concern, Barreca et al. consider a "donut RD" approach (see Section 2.2), through which they systematically drop observations with values close to 1500. While the authors confirm the results found in Almond et al., they report that estimates are highly sensitive to the exclusion of observations in the immediate neighborhood of the VLBW threshold.

Finally, we document the work of Bharadwaj et al. (2013) who use data from Chile and Norway to investigate the effects of improved early-life medical care on mortality and long-run academic achievement in school. Similar to the previous studies, the authors exploit the VLBW classification that assigns infants the treatment to special health care. The authors find a significant impact of extra medical care at birth on mortality rates and subsequent test scores in school.

### 4.5.2 | Health insurance

A longstanding question in economics and medicine is whether health insurance affects health outcomes. Some evidence is provided by Card et al. (2008), who for the United States, examine the impact of insurance coverage on health care utilization. More specifically, the authors evaluate a significant and discontinuous increase in insurance coverage at the 65 years of age threshold, generated by the rules of Medicare, the largest medical insurance program. Among their findings, the authors report an increase in the probability of routine doctor visits, and in procedures like bypass surgery and joint replacement. Later, Card et al. (2009) measure the health effects of Medicare eligibility for relatively sick patients, that is, people admitted through emergency rooms and with severe illnesses. The authors find a positive effect on the intensity of treatment for acutely ill patients with non-deferrable conditions and a reduction of patient mortality.

In general, evidence for developing economies is less abundant given the fewer resources available to subsidize health insurance programs. In these countries, most policies focus instead on improving access to health care services. One example is found in Bernal et al. (2017), which study the effects of large-scale health insurance access on various measures of health care utilization and out-of-pocket expenditures. Particularly, the authors center their investigation on the Peruvian public insurance program "*Seguro Integral de Salud*" targeted to poor individuals working in the informal labor market. The authors use a welfare index as a continuous forcing variable, where an individual who is not formally employed is eligible for free health insurance as long as her *Household Targeting Index* lies below a given threshold. As a result, Bernal et al. find large effects on several measures of curative care use, coupled with increases in out-of-pocket expenditures.

Finally, Camacho and Conover (2013) estimate the causal effect of health insurance provision on pregnant mothers and newborns' health quality. The authors use the *Poverty Index Score* as running variable, which again determines eligibility to subsidized health insurance. Camacho and Conover report a reduction of infants' low birth weight and better healthcare access to mothers.

### 4.5.3 | Retirement

There is a growing literature attempting to understand the effect of retirement on health outcomes, such as subjective and objective health measures, health behavior, and health care utilization. However, a fundamental challenge is that retirement can be endogenous, i.e. factors related to health can also affect a person's retirement decision. For example, Johnston and Lee (2009) argue that differences in discount rates can affect investment in health and also attachment to the labor force. Authors have often addressed this problem by using Statutory Retirement ages, as well as financial incentives as sources of discontinuous variation for identifying causal effects.

One example is Johnston and Lee (2009) who estimate the impact of retirement on health for a sample of English men. Specifically, the authors compare subjective and objective health measures for males below and above the 65-age retirement threshold. Interestingly, they omit individuals with a university degree claiming that degree holders are more likely to retire early (before the age of 65) from private pension funds. As a result, they find that retirement increases individual's self-reported sense of well-being and improvements in mental health, but not necessarily that of their physical health.

In turn, Eibich (2015) investigates the underlying mechanisms of retirement, centering on Germany as a case study. In contrast to the findings presented in Section 4.3.2, where health insurance coverage increased health care utilization, in this case the author finds a reduction in utilization which is linked to a significant improvement in health. That is, retirement offers relief from stress, the ability to exercise more frequently, and increased sleep duration. These factors have an impact on self-reported health, which the author further examines across different values of age, gender, education, occupation, and other sociodemographic characteristics.

Finally, we document Müller and Shaikh (2018) who present evidence on intra-household retirement externalities. In other words, the author explores the health effects of family members when an individual retires. Among their findings, Müller and Shaikh report that retiree's partners decrease physical activity and increases alcohol consumption. Hence, objective and subjective health are negatively affected by spousal retirement.

## 4.6 | Labor market

While education holds the title for most RDD studies in history, the area of labor is arguably the runner-up. As shown in Table 7, most of the research in this field seems to gravitate towards micro-labor, where financial incentives and (disincentives) affect the behavior of individuals and households. Typically, this concern has been addressed by examining the effects of extended benefits on unemployment duration and post-unemployment outcomes. An example of a discontinuity in this setting is the maximum duration of benefit entitlements. Besides, we document that a long-standing issue in the field relates to the effects of unionizing, and union representation.

Further, we recognize that some authors have studied the effects of reemployment services while others have looked at the effects of retirement. However, these designs mostly overlap with the field of health, therefore we will not detail those studies in this particular section.

We document that nearly 37% of the designs in labor use fuzzy discontinuities. For instance, Chen and Van der Klaauw (2008) evaluate the work-disincentive effects of the Disability Insurance (DI) program during the 1990s. Particularly, the authors find that the labor force participation rate of DI beneficiaries would have been at most 20 percentage points higher. Chen and Van der

**TABLE 7**  Common variables in labor market

| Relative frequency | Common outcomes | Common treatment and running variable | Fuzzy/Sharp |
|---|---|---|---|
| 18% | - Unemployment duration<br>- Post-unemployment variables (stability, wages, etc.) | Extended unemployment benefits ↔ Age | Sharp |
| 16% | - Labor supply<br>- Hours of home production | Retirement ↔ age | Fuzzy/Sharp |
| 5% | - Establishment closure<br>- Employment<br>- Output<br>- Wages | Union recognition ↔ vote share | Sharp |
| 11% | - Re-employment probability<br>- Earnings<br>- Mean weeks and UI benefits received | Reemployment services ↔ program criteria | Sharp |

Klaauw exploit the fact that the DI eligibility decision process is partly determined by an individual's age. However, DI also depends on other unobservable factors, such as residual functional capacity and illiteracy (education). Also, there are cases where rule-based recommendations are overruled by administrators, making the design even more fuzzy.

Another example is Stancanelli and Van Soest (2012) who for the French case study the causal effect of retirement on hours of home production of individuals and their partners. The identification strategy relies on the legislation on early retirement, which in France sets 60 years as the earliest retirement age for most workers. However, the fuzziness stems from people retiring earlier, due to special early retirement schemes and sector-specific agreements. It is also the case that some people retire well after their 60 years of age. Consequently, the probability jump of retirement is less than one at the 60-year threshold, but nonetheless greater than zero. Stancanelli and Van Soest find that there is a substantial increase in the hours of housework for both genders but not in symmetry. That is, the retirement of female partners significantly reduces the housework done by men, but not vice-versa.

### 4.6.1 | Financial incentives

Several authors have used discontinuities in unemployment benefits to assess the impact on unemployment duration. Regularly, benefits are a discontinuous function of age, however as will be reported below, authors have employed a number of techniques, even spatial RDDs to shed light on this issue. The main hypothesis is generally that extended benefits discourage searching for a job, and therefore, it leads to prolonged unemployment.

For instance, Lalive (2008) studies the link between labor market behavior and public transfers paid to the unemployed. More specifically, the author studies the effect of extending unemployment benefits from 30 to 209 weeks by exploiting a policy embedded in the Austrian unemployment insurance system. Namely, Austria temporarily implemented the *Regional Extended Benefit Program* (REBP) covering job seekers who entered unemployment at the age of 50 or older, and

who had been living in certain regions of Austria for at least 6 months. The study measures the discontinuity in unemployment duration at the assignment thresholds of age 50 and distance (in minutes) to the border. Lalive finds that the duration of job search is prolonged by at least 0.09 and 0.32 weeks (per additional week of benefits) among men and women, respectively.

Another example is presented in Card et al. (2007) who also for the Austrian case provide evidence of the excess sensitivity of job search behavior to disposable income. In essence, the authors employ discontinuities in the eligibility criteria for severance pay and extended unemployment insurance (UI) benefits, where individuals with only 3 or more years of job tenure were eligible. As such, they compare the search behavior of workers who were laid off just before and just after the cutoff for severance pay eligibility. Employing a very similar approach, the effects of extended benefits are identified (in this case the discontinuity is a function of the number of working months in the past 5 years). The authors find that the lump-sum severance payment equals 2 months of earnings, and that the extension of the potential duration of UI benefits lowers the job-finding rate. Also, they find that increases in job search have little or no effect on job match quality.

Lemieux and Milligan (2008) use both fuzzy and sharp designs to estimate the effect of higher social assistance benefits using a policy in the province of Quebec that paid much lower social assistance benefits to individuals without children and younger than 30 years. The authors find evidence that more generous social assistance reduces employment by 3–5 percentage points.

Finally, using data from Norway, Kostol and Mogstad (2014) analyze the effects of providing financial incentives to DI recipients to encourage them to return to work. To give some context, in 2005 the Norwegian government introduced a program that reduced the benefits of approximately $0.6 for each $1 in earnings that the DI recipients accumulated above a substantial gainful activity threshold. However, only recipients who had been awarded DI before January 1, 2004 were eligible for the return-to-work program. Thus, the recipients awarded DI just before (after) January 1 of 2004 are defined as the treatment (control) group. The authors find that many DI recipients have considerable capacity to work, which can be effectively induced by providing financial work incentives.

### 4.6.2 | Unionization

We next turn our focus to studies that evaluate the economic effects of unionization on the labor market. Here, the main identification challenge is that unionization does not occur at random. Instead, the correlation between union status and employer outcomes (e.g., employment, output) may introduce upward or downward biases to the true effect of unions. As DiNardo and Lee (2004) (p. 1385) state, "*unions may tend to organize at highly successful enterprises that are more likely to survive and, a union may be more likely to succeed when a firm is poorly managed or has faced recent difficulties*". To overcome these biases, several authors have turned to quasi-experiments inherent in union representation elections. For example, in the United States, most establishments become unionized as a partial consequence of a secret ballot election among workers. Thus, "close elections" are used to draw causal inference.

Using data from U.S. establishments during 1983–1999, DiNardo and Lee (2002) explore the causal effect of unionization on employer closure. The authors compare outcomes between establishments that faced elections where the union barely won, by one vote, and those that barely lost, also by one vote. By law, a simple majority vote in favor of the union requires the firm's

management to recognize and bargain with the victorious union in collective bargaining negotiations. However, DiNardo and Lee find no evidence of union effects on short and long-term employer survival rates. Employing the same approach DiNardo and Lee (2004) study the effects of unionization on business failures, employment, output, productivity, and wages. Once again, the authors exploit the discontinuous function between vote share and winning the union representation. This time the authors find a small effect on all outcomes (albeit estimates for wages are close to zero).

Finally, Lee and Mas (2012) estimate the effect of new private-sector unionization on publicly traded firm equity value for the period of 1961–1999. More precisely, the authors compare the stock market impact of close union election wins against close losses. The authors' findings point towards little evidence of a significant discontinuous relationship between the vote share and market returns.

### 4.6.3 | Reemployment services

Using data from Kentucky, Black et al. (2003) study the effect of the *Worker Profiling and Reemployment Service system*, WPRS. This program 'profiles' unemployment insurance (UI) claimants based on the predicted length of their unemployment spell or the predicted probability that they will consume their UI benefits. Technically, the requirement to receive reemployment services is allocated by profiling score up to capacity. And, within the marginal profiling score, random assignment allocates the mandatory services. The authors find that the program reduces the mean weeks of UI benefit receipt by near 2.2 weeks, reduces mean UI benefit received by about $143, and increases the subsequent earnings by over $1050.

Finally, de Giorgi (2005) investigate the effects of enhancing the (re)employment probability of the New Deal for Young People (NDYP), a major welfare-to-work policy in the UK. The author exploits the fact that participation in the program is compulsory and established by a deterministic rule. The policy is administered to everyone in the UK with 6 months of *Job Seeker Allowance*, plus being younger than 25 years of age. In this context, the treatment consists of job search assistance, training/education, subsidies, and job experience. The author finds that the program enhances employability by about 6%–7%.

## 4.7 | **Political economy**

In the field of political economy, most of the research centers on incumbency advantage (29% as shown in Table 8), particularly on the probability of being elected during the next political race. Another part of the literature has focused on issues related to public spending, where treatments include being an incumbent, partisan representation, political regime, and an increase in elected representatives. Popular running variables for these include voting share and population size. For example, Ferreira and Gyourko (2009) and Höhmann (2017) both investigate issues related to public spending, but they use initial vote share and jurisdictions' population size, respectively, as running variables.

Notably, in Table 1 we observe that almost 77% of the literature employs a sharp design. Since voting share is most common among running variables, it is almost always the case—by constitutional norm—that a voting majority (be it simple or absolute) perfectly determines the winner.[10]

**TABLE 8** Common variables in political economy

| Relative frequency | Common outcomes | Common treatment and running variable | Fuzzy/Sharp |
|---|---|---|---|
| 29% | - Probability of running<br>- Probability of winning next election<br>- Vote share in next election | Incumbency ↔ Initial vote share | Sharp |
| 26% | Expenditure or public spending | Increase in elected representatives ↔ Population size | Fuzzy/Sharp |
| | | Incumbency ↔ Initial vote share | Sharp |
| | | Partisan representation ↔ Vote share in election | Sharp |
| | | Local choice of political regime ↔ Population size | Fuzzy |
| 6% | Roll call votes | Incumbency ↔ Initial vote share | Sharp |

Nonetheless, some exceptions exist in the field where studies turn to a fuzzy framework. One example is Fiva and Halse (2016) who evaluate whether members of the Norwegian council use their position to lobby (i.e., obtain more public investment) for their hometowns. In principle, a switch from a left to a right-wing governor would adversely affect municipalities with left-wing representatives. However, the fuzziness stems from the fact that "regional alignment" is not entirely exogenous. That is, municipalities vary in both number and political affiliation of council representatives (e.g., they can have representatives from both parties or from none). In some cases, municipalities with left-wing representatives can ultimately align to the right. As such, the authors instrument political alignment with the cross-term between regional majority and the political party of each representative.[11] They find that public investment funds increase when the regional council is aligned with the choice of governor.

Using data from Finland and Sweden, Pettersson-Lidbom (2012) estimate the causal effect of council size on government spending for jurisdictions with an identical electoral system. The national laws regulating the council size in Finnish and Swedish local governments provide a source of exogenous variation in legislature size. However, while in Finland the council size is a deterministic function of population size, in Sweden the council's size is partly determined by statutory law, and hence the authors correct treatment compliance with a fuzzy design.

In turn, Hinnerich and Pettersson-Lidbom (2014) evaluate Sweden's transition from a non-democracy to a democratic electoral system in 1919 to compare how two types of political regimes (direct and representative democracies) redistribute income within relatively poor segments of the population.[12] Namely, local governments were required by the Swedish Local Government Act to have a representative democracy if their population size was larger than 1500, but they were free to choose between representative or direct democracy below this threshold. Using a fuzzy RDD, the authors find that direct democracies spend 40%–60% less on public welfare.

Another example employing a fuzzy design is found in Freier and Thomasius (2016) who investigate the importance of politicians' qualification (in terms of education and experience) on fiscal outcomes. Specifically, the authors consider cases in which the winner and runner-up differ in levels of education. The margin of victory ($m$) is defined as the difference in vote

**TABLE 9** Glossary of terms

| | |
|---|---|
| Ex-post facto experiment: | A quasi-experiment where individuals (or measurable units) are not randomly assigned. With the aid of tools and techniques, treatment and control groups can be comparable at baseline. |
| Running variable: | Variable used to sort and deliver treatment (also referred to as the *assignment variable*). |
| Sharp designs: | Designs where the rule-based mechanism perfectly (and deterministically) assigns treatment. |
| Fuzzy designs: | Similar to sharp designs but allowing for non-compliers (e.g., control and treatment crossovers). |
| Potential outcomes: | Hypothetical outcomes, for example, what would happen if an individual receives treatment, regardless if she actually received treatment. |
| Endogeneity: | Mostly refers to missing information (self-selection or omitted variables) that is useful in explaining the true effect of treatment (other endogeneity problems include simultaneity and measurement error). |
| Bandwidths: | Window size applied to the running variable which determines which observations are included in the analysis. The trade-off between bias and precision is a fundamental feature of bandwidths: larger bandwidths yield more precise, but potentially biased estimates. |
| Kernel regressions: | Non-parametric technique used to estimate non-linear associations. Kernels (and bandwidths) give higher weight (or restrict) observations that fall close to the mean. |
| Local Polynomials: | Functions used to capture the relationship between the running variable and outcome variable, close to the cutoff. While parametric in nature, the formal properties are similar to kernel methods (although more sensitive to observations far away from the cutoff point). The order of the polynomial is usually chosen according to the Akaike's criterion (penalized cross-validation). |
| Parametric estimation: | Estimations that assume probability distributions for the data. |
| Overlap assumption: | Assumption that guarantees that similar individuals are both treated and untreated. |
| Local projections: | Approximations conducted locally for each forecast horizon of interest. |

share between the more highly educated ($v_h$) and lower educated ($v_l$) candidate. Thus, non-negative values ($m = v_h - v_l \geq 0$) determine that a higher educated candidate gets into office. However, education is not presented as a binary variable, and rather captures the (continuous) probability of having a university degree.[13] Using a fuzzy design, the authors find that mayors with prior office experience tend to reduce local public debt, lower municipal expenditures, and decrease local taxes. In contrast, the authors find no significant effects of education levels on fiscal outcomes.

### 4.7.1 | Incumbency effects

In principle, winning an election can have a positive effect on the probability of being reelected. And, while incumbency advantages have been observed in a variety of contexts, few studies have actually been able to quantify their effect. Using data from the United States House of Representa-

tives, Lee (2008) was the first in estimating the incumbency advantage by exploiting the inherent voting discontinuity within Congressional elections.[14] Specifically, he compares the subsequent electoral outcomes of candidates (and that of their parties) that barely won to those that barely lost. Essentially, as one compares closer and closer electoral races, candidates are ex-ante comparable in all ways other than in their eventual incumbency status.

Lee defines the running variable as the Democratic vote share margin of victory. That is, the Democratic vote share minus the share of its strongest opponent, virtually always a Republican, with a cutoff at 0%. Among his findings, he shows that party incumbency increases the probability that a political party will retain the districts' seat by 0.40–0.45 and that losing an election reduces the probability of a candidate running again for office by 0.43.

Relatedly, Lee et al. (2004) investigate whether parties benefiting from an incumbency advantage change or stick to their political views. Namely, electoral competition induces candidates to move towards the center and compromise, so in this sense voters "affect policies". On the other hand, it is difficult for politicians to make a credible move towards moderation, so voters favor candidates with more fixed positions, and thus "elect policies". The authors estimate the effect of a candidate's electoral strength on subsequent roll-call voting records and find that the degree of electoral strength has no effect on legislators' voting behavior. Candidates with weak electoral support do not adopt more moderate positions than stronger candidates, suggesting that voters do not affect policies, instead, they appear to merely elect them.

Albouy (2011) arrives to similar results to those in Lee et al. (2004), but notes that "*junior members of Congress prefer to vote more extremely than senior members, independently of their electoral strength*" (p. 162). When correcting for this bias he finds that candidates in fact moderate their policy choices in response to electoral competition, that is, senators vote more moderately in the years prior to their next election.

More on the effects of incumbency can be found in Broockman (2009), Uppal (2009), Freier (2015), and Fiva and Røhr (2018).

### 4.7.2 | Government size and public spending

Several authors have examined the effect of diverse treatments on public spending. For example, in the context of mayoral elections, Ferreira and Gyourko (2009) evaluate whether partisan political differences at the local level in the United States affect outcome variables such as government size, allocation of public spending, and crime rates. More specifically, the authors compare cities where democratic candidates barely won an election with cities where they barely lost. Similar to Lee (2008), they use the margin of victory as running variable. While the authors find no evidence of a strong partisan influence on the aforementioned outcome variables, they do report a large incumbency advantage in winning the next election. The authors attribute their results to a high degree of household homogeneity that potentially encourages local politicians to credibly commit to moderate policies (avoiding strategic extremism).

Similarly, Pettersson-Lidbom (2008) estimates the causal effect of party control on fiscal and economic policies using a panel data set from the Swedish local governments. The author's results show that left-wing governments spend and tax 2%–3% more than right-wing governments. Also, left-wing governments have 7% lower unemployment rates, partly explained by employing 4% more workers than right-wing governments.

Another example is given by Albouy (2013) who investigates how party membership of legislators in the United States Congress may influence the distribution of federal funds. Again the

author exploits data from elections that are won by close margins. Albouy finds that if a state sends a Senate delegation with both members in the majority, it receives 2% more of government grants. Further, states represented by republicans receive more spending for defense and transportation, while states represented by democrats receive more funding earmarked for education and urban development.

In the related literature, some authors use population size as running variable when evaluating the effects on expenditure. Such is the case of Pettersson-Lidbom (2012) who estimates the effect of council size on government spending for jurisdictions with an identical electoral system, in his example Finland and Sweden. He finds a negative effect, that is, the larger the size of the legislature, the smaller the size of government, and argues that "*more legislators can better control a budget maximizing bureaucracy*".

In the context of German municipal councils, Höhmann (2017) also provides evidence of the effect of legislature size (increase in the number of representatives) on public spending. In essence, the municipal council size is a discontinuous function of a jurisdictions' population size. Höhmann also reports a negative effect of legislature size on public spending.

### 4.7.3 | Taxes and fees

Berger et al. (2016) study the link between taxes and evasion. The authors approximate the problem by studying TV license fees in Austria. Austria, as in many other countries, has mandatory TV and radio license fees to finance public broadcasting. Berger et al. exploit the discontinuous change in fees at state borders and implement a spatial RDD by using the distance to the closest border as assignment variable. They follow Lalive (2008) in computing the distance from the border by the driving time (instead of the standard Euclidean distance) since driving time better reflects the Austrian topography. Hence, municipalities with a negative [positive] distance hold lower [higher] fees, with a threshold of 0:00 at the border. Using parametric and non-parametric approaches, the authors find a positive effect of fees on evasion. On average, they find an 18% fee increase at state borders accompanied by an increase in the evasion rate of 5 percentage points. A back-of-the-envelope calculation indicates that a 1% increase in fees raises evasion by 0.3 percentage points.

By comparison, Luechinger and Roth (2016) estimate the effect of a mileage tax for heavy vehicle on traffic volume. In 2001, Switzerland introduced a distance-related fee on vehicles weighing more than 3.5 tons. Notably, the authors evaluate changes around the policy introduction date. Thus, the RDD uses time as running variable.[15] The authors find that mileage taxes reduce truck traffic by 4%–6%. They find no significant effects on car traffic which the authors claim as suggestive evidence for an increase in rail freight traffic. Finally, they find a reduction of nitrogen oxide pollution of 5.6% but only in close proximity to roads.

## 5 | CONCLUDING REMARKS

In this survey, we bring together the entire RDD literature over the past 60 years. Namely, we document its origins in the 1960s and the various formalization waves henceforth (in the 1970s and 2000s).

Particularly, we break the literature down by economic field, highlighting the main outcomes, treatments, and running variables employed. Overall, we see some topics in economics gaining

importance through time, like the cases of health, finance, crime, environment, and political economy. In particular, we highlight applications in finance as the most novel. Nonetheless, education stands out as the uncontested RDD champion through time, with the field of labor as runner-up.

Finally, we dissect the literature into fuzzy and sharp designs. Specifically, we provide some intuition as to why some criteria produced non-compliers and also report the historical distribution of fuzzy versus sharp designs.

## ORCID

*Mauricio Villamizar-Villegas* 🄳 https://orcid.org/0000-0001-8866-9638

## ENDNOTES

[1] Harold Hotelling in "The Collected Economics Articles of Harold Hotelling" (p. 52). Mary C. Bateson in "Composing a Life" (p. 13).

[2] This is also the case for studies like Cook (2008) which mainly focuses on papers related to psychology, education, and statistics, De la Cuesta and Imai (2016) which cover 20 studies, Venkataramani et al. (2016) which focus on health-related studies, and Valencia (2020) which covers articles only for the top journals in economics and economic history.

[3] A more recent version of the McCrary test is found in Cattaneo et al. (2018) with improved size and power properties based on density discontinuity. The major contribution of the authors is the development of an intuitive and easy-to-implement nonparametric density estimator based on local polynomial techniques, which does not require pre-binning or any other transformation of the data while still being fully boundary adaptive and automatic.

[4] For further reading, studies that consider the possibility of other discontinuous jumps at the same cutoff include Imbens (2004), Lee et al. and Battistin and Rettore (2008).

[5] Villamizar-Villegas and Önder (2020) extend the study of heterogeneous treatment effects to a time series setting. Also, Bertanha (2020) proposes new estimation and inference methods for global average effects in a setting with many thresholds.

[6] Other relevant studies include: Lee and Card (2008), Yu (2012), Yanagi (2014), Barreca et al. (2016), and Pei and Shen (2017).

[7] Note that $E[D_i|x_i] = 1 \cdot Pr[D_i = 1|x_i] + 0 \cdot Pr[D_i = 0|x_i]$.

[8] These methods are originally proposed in Spiegelman (1976), Spiegelman (1977), Maddala and Lee (1976) and Barnow et al. (1978).

[9] The number of neighbors varies depending on the width of the moving average window.

[10] See for example, Lee et al. (2004), Lee (2008), Eggers and Hainmueller (2009), Ferreira and Gyourko (2009), Firpo et al. (2015), and Freier (2015).

[11] According to Fiva and Halse, these cross-terms captures the municipalities' probability of becoming politically aligned when a governor takes office.

[12] Representative democracies held regular elections every 4 years, where citizens vote for political parties. In contrast, direct democracies congregate citizens at town meetings at least three times per year.

[13] For some candidates, the education variable takes the value of the average share of graduates among employees in the Bavaria region of Germany.

[14] Lee first published his work as a working paper in 2001 (Lee, 2001).

[15] Luechinger and Roth include time polynomials to control for observable and unobservable factors that might affect outcomes over time.

## REFERENCES

Aiken, L. S., West, S. G., Schwalm, D. E., Carroll, J. L., & Hsiung, S. (1998). Comparison of a randomized and two quasi-experimental designs in a single outcome evaluation: Efficacy of a university-level remedial writing program. *Evaluation Review*, *22*(2), 207–244.

Albouy, D. (2011). Do voters affect or elect policies? A new perspective, with evidence from the us senate. *Electoral Studies*, *30*(1), 162–173.

Albouy, D. (2013). Partisan representation in congress and the geographic distribution of federal funds. *The Review of Economics and Statistics*, *95*(1), 127–141.

Almond, D., Doyle Jr, J. J., Kowalski, A. E., & Williams, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics*, *125*(2), 591–634.

Anderson, K., Gong, X., Hong, K., & Zhang, X. (2016). Do selective high schools improve student achievement? Effects of exam schools in China. *China Economic Review*, *40*(C), 121–134.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, *15*(4), 69–85.

Angrist, J. D., & Lavy, V. (1999). Using Maimonides' rule to estimate the effect of class size on scholastic achievement. *The Quarterly Journal of Economics*, *114*(2), 533–575.

Angrist, J. D., & Pischke, J.-S. (2009). *Mostly harmless econometrics: An empiricist's companion*. Princeton University Press.

Asadullah, M. N. (2005). The effect of class size on student achievement: Evidence from Bangladesh. *Applied Economics Letters*, *12*(4), 217–221.

Aung, N., Nguyen, H.-T.-M., & Sparrow, R. (2019). The impact of credit policy on rice production in Myanmar. *Journal of Agricultural Economics*, *70*(2), 426–451.

Balthrop, A. T., & Schnier, K. E. (2016). A regression discontinuity approach to measuring the effectiveness of oil and natural gas regulation to address the common-pool externality. *Resource and Energy Economics*, *44*(C), 118–138.

Barnow, B. S. (1972). Conditions for the presence or absence of a bias in treatment effect: Some statistical models for head start evaluation. Discussion papers 122-72, Institute for Research on Poverty, University of Wisconsin-Madison.

Barnow, B. S., Cain, G. G., & Goldberger, A. S. (1978). Issues in the analysis of selection bias. *Unpublished paper, August*.

Barreca, A. I., Guldi, M., Lindo, J. M., & Waddell, G. R. (2011). Saving babies? Revisiting the effect of very low birth weight classification. *The Quarterly Journal of Economics*, *126*(4), 2117.

Barreca, A. I., Lindo, J. M., & Waddell, G. R. (2016). Heaping-induced bias in regression-discontinuity designs. *Economic Inquiry*, *54*(1), 268–293.

Barrera, F., Linden, L. L., & Urquiola, M. (2013). The effects of user fee reductions on enrollment: Evidence from a quasi-experiment. *Manuscript. Department of Economics. The University of Texas at Austin*.

Bartalotti, O., Brummet, Q., & Dieterle, S. (2021). A correction for regression discontinuity designs with group-specific mismeasurement of the running variable. *Journal of Business & Economic Statistics*, *39*(3):1–16.

Battistin, E., & Rettore, E. (2008). Ineligibles and eligible non-participants as a double comparison group in regression-discontinuity designs. *Journal of Econometrics*, *142*(2), 715–730.

Becker, G. S. (1968). Crime and punishment: An economic approach. *Journal of Political Economy*, *76*, 169.

Berger, M., Fellner-Röhling, G., Sausgruber, R., & Traxler, C. (2016). Higher taxes, more evasion? Evidence from border differentials in tv license fees. *Journal of Public Economics*, *135*, 74–86.

Berk, R. A., & de Leeuw, J. (1999). An evaluation of California's inmate classification system using a generalized regression discontinuity design. *Journal of the American Statistical Association*, *94*(448), 1045–1052.

Berk, R. A., & Rauma, D. (1983). Capitalizing on nonrandom assignment to treatments: A regression-discontinuity evaluation of a crime-control program. *Journal of the American Statistical Association*, *78*(381), 21–27.

Bernal, N., Carpio, M. A., & Klein, T. J. (2017). The effects of access to health insurance: Evidence from a regression discontinuity design in Peru. *Journal of Public Economics*, *154*, 122–136.

Bertanha, M. (2020). Regression discontinuity design with many thresholds. *Journal of Econometrics*, *218*(1), 216–241.

Bertanha, M., & Imbens, G. W. (2020). External validity in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, *38*(3), 593–612.

Bharadwaj, P., Jack, W., & Suri, T. (2019). Fintech and household resilience to shocks: Evidence from digital loans in Kenya. Technical report. National Bureau of Economic Research.

Bharadwaj, P., Løken, K. V., & Neilson, C. (2013). Early life health interventions and academic achievement. *American Economic Review*, *103*(5), 1862–1891.

Black, D. A., Galdo, J., & Smith, J. A. (2005). Evaluating the regression discontinuity designusing experimental data. *Unpublished manuscript*.

Black, D. A., Smith, J. A., Berger, M. C., & Noel, B. J. (2003). Is the threat of reemployment services more effective than the services themselves? Evidence from random assignment in the UI system. *American Economic Review*, *93*(4), 1313–1327.

Black, S. E. (1999). Do better schools matter? Parental valuation of elementary education. *The Quarterly Journal of Economics*, *114*(2), 577–599.

Bonaldi, P., & Villamizar-Villegas, M. (2018). An auction-based test of private information in an interdealer FX market. Borradores de Economia 1049, Banco de la Republica de Colombia.

Broockman, D. E. (2009). Do Congressional Candidates Have Reverse Coattails? Evidence from a regression discontinuity design. *Political Analysis*, *17*(04), 418–434.

Calonico, S., Cattaneo, M. D., & Farrell, M. H. (2020). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal*, *23*(2), 192–210.

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

Camacho, A., & Conover, E. (2013). Effects of subsidized health insurance on newborn health in a developing country. *Economic Development and Cultural Change*, *61*(3), 633–658.

Campbell, D. T. (1969). Reforms as experiments. *American Psychologist*, *24*(4), 409.

Campbell, D. T., & Erlebacher, A. (1970). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. *Compensatory Education: A National Debate*, *3*, 185–210.

Campbell, D. T., & Stanley, J. C. (1963). Experimental and quasi-experimental designs for research. In *Handbook of research on teaching* (pp. 1–84). Boston, MA: Houghton, Mifflin and Company.

Canton, E. & Blom, A. (2004). Can student loans improve accessibility to higher education and student performance? An impact study of the case of SOFES, Mexico. Policy Research Working Paper Series 3425, The World Bank.

Card, D., Chetty, R., & Weber, A. (2007). Cash-on-hand and competing models of intertemporal behavior: New evidence from the labor market. *The Quarterly Journal of Economics*, *122*(4), 1511–1560.

Card, D., Dobkin, C., & Maestas, N. (2008). The impact of nearly universal insurance coverage on health care utilization: Evidence from medicare. *American Economic Review*, *98*(5), 2242–58.

Card, D., Dobkin, C., & Maestas, N. (2009). Does medicare save lives? *The Quarterly Journal of Economics*, *124*(2), 597–636.

Card, D., & Krueger, A. (1994). Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, *84*(4), 772–93.

Cascio, E. U., & Lewis, E. G. (2006). Schooling and the armed forces qualifying test evidence from school-entry laws. *Journal of Human resources*, *41*(2), 294–318.

Cattaneo, M. D., Idrobo, N., & Titiunik, R. (2019). *A practical introduction to regression discontinuity designs: Foundations*. Cambridge University Press.

Cattaneo, M. D., Jansson, M., & Ma, X. (2018). Manipulation testing based on density discontinuity. *The Stata Journal*, *18*(1), 234–261.

Cattaneo, M. D., Titiunik, R., Vazquez-Bare, G., & Keele, L. (2016). Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, *78*(4), 1229–1248.

Cellini, S. R., Ferreira, F., & Rothstein, J. (2010). The value of school facility investments: Evidence from a dynamic regression discontinuity design. *The Quarterly Journal of Economics*, *125*(1), 215.

Chang, Y.-C., Hong, H., & Liskovich, I. (2014). Regression discontinuity and the price effects of stock market indexing. *The Review of Financial Studies*, *28*(1), 212–246.

Chaplin, D. D., Cook, T. D., Zurovac, J., Coopersmith, J. S., Finucane, M. M., Vollmer, L. N., & 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.

Chava, S., & Roberts, M. R. (2008). How does financing impact investment? The role of debt covenants. *The Journal of Finance*, *63*(5), 2085–2121.

Chay, K. Y., Dobkin, C., & Greenstone, M. (2003). The clean air act of 1970 and adult mortality. *Journal of Risk and Uncertainty*, *27*(3), 279–300.

Chay, K. Y., & Greenstone, M. (2003). Air quality, infant mortality, and the Clean Air Act of 1970. NBER Working Papers 10053, National Bureau of Economic Research, Inc.

Chay, K. Y., & Greenstone, M. (2005). Does air quality matter? Evidence from the housing market. *Journal of Political Economy*, *113*(2), 376–424.

Chay, K. Y., McEwan, P. J., & Urquiola, M. (2005). The central role of noise in evaluating interventions that use test scores to rank schools. *American Economic Review*, *95*(4), 1237–1258.

Chen, M. K., & Shapiro, J. M. (2003). Does prison harden inmates? a discontinuity-based approach." Cowles foundation discussion papers 1450, cowles foundation for research in economics, Yale University.

Chen, S., & Van der Klaauw, W. (2008). The work disincentive effects of the disability insurance program in the 1990s. *Journal of Econometrics*, *142*(2), 757–784.

Chen, T., Huang, Y., Lin, C., & Sheng, Z. (2020). Finance and firm volatility: Evidence from small business lending in China. *Management Science*, 0(0).

Chen, Y., Ebenstein, A., Greenstone, M., & Li, H. (2013). Evidence on the impact of sustained exposure to air pollution on life expectancy from China's Huai river policy. *Proceedings of the National Academy of Sciences*, *110*(32), 12936–12941.

Chiang, H. D., Hsu, Y.-C., & Sasaki, Y. (2019). Robust uniform inference for quantile treatment effects in regression discontinuity designs. *Journal of Econometrics*, *211*(2), 589–618.

Chiang, H. D., Kato, K., Sasaki, Y., & Ura, T. (2021). Linear programming approach to nonparametric inference under shape restrictions: With an application to regression kink designs. Papers 2102.06586, arXiv.org.

Clark, D., & Royer, H. (2013). The effect of education on adult mortality and health: Evidence from Britain. *American Economic Review*, *103*(6), 2087–2120.

Cook, P. J., & Kang, S. (2016). Birthdays, schooling, and crime: Regression-discontinuity analysis of school performance, delinquency, dropout, and crime initiation. *American Economic Journal: Applied Economics*, *8*(1), 33–57.

Cook, T. D. (2008). 'Waiting for life to arrive': A history of the regression-discontinuity design in psychology, statistics and economics. *Journal of Econometrics*, *142*(2), 636–654.

Cook, T. D., & Wong, V. C. (2008). Empirical tests of the validity of the regression discontinuity design: Implications for its theory and its use in research practice. *Annals of Economics and Statistics*, (91–92), 127–150.

Cumming, D., Farag, H., Johan, S., & McGowan, D. (2021). The digital credit divide: Marketplace lending and entrepreneurship. *Journal of Financial and Quantitative Analysis*, forthcoming.

Dang, T. (2018). Do the more educated utilize more health care services? Evidence from Vietnam using a regression discontinuity design. *International Journal of Health Economics and Management*, *18*(3), 277–299.

Datar, A. (2006). Does delaying kindergarten entrance give children a head start? *Economics of Education Review*, *25*(1), 43–62.

Davezies, L., & Le Barbanchon, T. (2017). Regression discontinuity design with continuous measurement error in the running variable. *Journal of Econometrics*, *200*(2), 260–281.

Davis, L. W. (2008). The effect of driving restrictions on air quality in Mexico City. *Journal of Political Economy*, *116*(1), 38–81.

de Giorgi, G. (2005). Long-term effects of a mandatory multistage program: The New Deal for Young People in the UK. IFS Working Papers W05/08, Institute for Fiscal Studies.

De la Cuesta, B., & Imai, K. (2016). Misunderstandings about the regression discontinuity design in the study of close elections. *Annual Review of Political Science*, *19*, 375–396.

Magalhaes, L., Hangartner, D., Hirvonen, S., Meriläinen, J., Ruiz, N., & Tukiainen, J. (2020). How much should we trust regression discontinuity design estimates? Evidence from experimental benchmarks of the incumbency advantage.

Dee, T., & Lan, X. (2015). The achievement and course-taking effects of magnet schools: Regression-discontinuity evidence from urban China. *Economics of Education Review*, *47*(C), 128–142.

Dell, M. (2010). The persistent effects of Peru's mining mita. *Econometrica*, *78*(6), 1863–1903.

Depew, B., & Eren, O. (2016). Born on the wrong day? School entry age and juvenile crime. *Journal of Urban Economics*, *96*, 73–90.

DiNardo, J., & Lee, D. S. (2002). The impact of unionization on establishment closure: A regression discontinuity analysis of representation elections. NBER Working Papers 8993, National Bureau of Economic Research, Inc.

DiNardo, J., & Lee, D. S. (2004). Economic impacts of new unionization on private sector employers: 1984-2001. NBER Working Papers 10598, National Bureau of Economic Research, Inc.

DiNardo, J., & Lee, D. S. (2011). Program evaluation and research designs. In *Handbook of labor economics* (vol. 4, pp. 463–536). Elsevier.

Dong, Y. (2015). Regression discontinuity applications with rounding errors in the running variable. *Journal of Applied Econometrics*, *30*(3), 422–446.

Dong, Y. (2019). Regression discontinuity designs with sample selection. *Journal of Business & Economic Statistics*, *37*(1), 171–186.

Dykstra, S., Glassman, A., Kenny, C., & Sandefur, J. (2019). Regression discontinuity analysis of gavi's impact on vaccination rates. *Journal of Development Economics*, *140*, 12–25.

Eggers, A. C., Freier, R., Grembi, V., & Nannicini, T. (2018). Regression discontinuity designs based on population thresholds: Pitfalls and solutions. *American Journal of Political Science*, *62*(1), 210–229.

Eggers, A. C., & Hainmueller, J. (2009). Mps for sale? Returns to office in postwar Britishpolitics. *American Political Science Review*, *103*(4), 513–533.

Eibich, P. (2015). Understanding the effect of retirement on health: Mechanisms and heterogeneity. *EconStor Open Access Articles*, 1–12.

Falato, A., & Liang, N. (2016). Do creditor rights increase employment risk? Evidence from loan covenants. *The Journal of Finance*, *71*(6), 2545–2590.

Fan, J., & Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66* (vol. 66). CRC Press.

Feir, D., Lemieux, T., & Marmer, V. (2016). Weak identification in fuzzy regression discontinuity designs. *Journal of Business & Economic Statistics*, *34*(2), 185–196.

Ferreira, F., & Gyourko, J. (2009). Do political parties matter? Evidence from us cities. *The Quarterly Journal of Economics*, *124*(1), 399–422.

Filmer, D., & Schady, N. (2009). School enrollment, selection and test scores. Policy Research Working Paper Series 4998, The World Bank.

Firpo, S., Ponczek, V., & Sanfelice, V. (2015). The relationship between federal budget amendments and local electoral power. *Journal of Development Economics*, *116*, 186–198.

Fiva, J. H., & Halse, A. H. (2016). Local favoritism in at-large proportional representation systems. *Journal of Public Economics*, *143*, 15–26.

Fiva, J. H., & Røhr, H. L. (2018). Climbing the ranks: Incumbency effects in party-list systems. *European Economic Review*, *101*(C), 142–156.

Foote, A., Schulkind, L., & Shapiro, T. M. (2015). Missed signals: The effect of ACT college-readiness measures on post-secondary decisions. *Economics of Education Review*, *46*(C), 39–51.

Freier, R. (2015). The mayor's advantage: Causal evidence on incumbency effects in German mayoral elections. *European Journal of Political Economy*, *40*(PA), 16–30.

Freier, R., & Thomasius, S. (2016). Voters prefer more qualified mayors, but does it matter for public finances? Evidence for germany. *International Tax and Public Finance*, *23*(5), 875–910.

Frölich, M., & Huber, M. (2018). Including covariates in the regression discontinuity design. *Journal of Business & Economic Statistics*, *37*, 736–748.

Frölich, M., & Melly, B. (2008). Quantile treatment effects in the regression discontinuity design. IZA Discussion Papers 3638, Institute of Labor Economics (IZA).

Fu, S., & Gu, Y. (2017). Highway toll and air pollution: Evidence from Chinese cities. *Journal of Environmental Economics and Management*, *83*(C), 32–49.

Garmaise, M. J. (2013). The attractions and perils of flexible mortgage lending. *Review of Financial Studies*, *26*(10), 2548–2582.

Garmaise, M. J., & Natividad, G. (2017). Consumer default, credit reporting, and borrowing constraints. *The Journal of Finance*, *72*(5), 2331–2368.

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

Goldberger, A. S. (1972). Selection bias in evaluating treatment effects: Some formal illustrations. Discussion Paper 123-72, Institute for Research on Poverty. Madison, WI.

Goodman, J. (2008). Who merits financial aid?: Massachusetts' Adams scholarship. *Journal of Public Economics*, *92*(10-11), 2121–2131.

Goodman, J., Melkers, J., & Pallais, A. (2019). Can online delivery increase access to education? *Journal of Labor Economics*, *37*(1), 1–34.

Greenstone, M., & Gallagher, J. (2008). Does hazardous waste matter? Evidence from the housing market and the superfund program. *The Quarterly Journal of Economics*, *123*(3), 951–1003.

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

Gu, Y., Mao, C. X., & Tian, X. (2017). Banks' interventions and firms' innovation: Evidence from debt covenant violations. *The Journal of Law and Economics*, *60*(4), 637–671.

Guryan, J. (2001). Does money matter? Regression-discontinuity estimates from education finance reform in Massachusetts. NBER Working Papers 8269, National Bureau of Economic Research, Inc.

Hahn, J., Todd, P., & Van der Klaauw, W. (1999). Evaluating the effect of an antidiscrimination law using a regression-discontinuity design. NBER Working Papers 7131, National Bureau of Economic Research, Inc.

Hahn, J., Todd, P., & Van der Klaauw, W. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, *69*(1), 201–09.

Hausman, C., & Rapson, D. S. (2018). Regression discontinuity in time: Considerations for empirical applications. *Annual Review of Resource Economics*, *10*(1), 533–552.

Heckman, J. J., Lalonde, R. J., & Smith, J. A. (1999). The economics and econometrics of active labor market programs. In O. Ashenfelter & D. Card (Eds.), *Handbook of labor economics*. Elsevier 3(31), 1865–2097.

Hinnerich, B. T., & Pettersson-Lidbom, P. (2014). DeDemocracy, redistribution, and political participation: Evidence from Sweden 1919-1938. *Econometrica*, *82*(3), 961–993.

Hjalmarsson, R. (2009). Juvenile jails: A path to the straight and narrow or to hardened criminality? *Journal of Law and Economics*, *52*(4), 779–809.

Höhmann, D. (2017). The effect of legislature size on public spending: Evidence from a regression discontinuity design. *Public Choice*, *173*(3), 345–367.

Holmes, T. J. (1998). The effect of state policies on the location of manufacturing: Evidence from state borders. *Journal of Political Economy*, *106*(4), 667–705.

Hong, K., & Zimmer, R. (2016). Does investing in school capital infrastructure improve student achievement? *Economics of Education Review*, *53*(C), 143–158.

Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. *The Quarterly Journal of Economics*, *115*(4), 1239–1285.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics*, *86*(1), 4–29

Imbens, G. W., & Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *Review of Economic Studies*, *79*, 933–959.

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

Jacob, B. A., & Lefgren, L. (2004a). The impact of teacher training on student achievement quasi-experimental evidence from school reform efforts in Chicago. *Journal of Human Resources*, *39*(1), 50–79.

Jacob, B. A., & Lefgren, L. (2004b). Remedial education and student achievement: A regression-discontinuity analysis. *The Review of Economics and Statistics*, *86*(1), 226–244.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. *Journal of Financial Economics*, *3*(4), 305–360.

Johnston, D. W., & Lee, W.-S. (2009). Retiring to the good life? the short-term effects of retirement on health. *Economics Letters*, *103*(1), 8–11.

Jordá, Ó. (2005). Estimation and inference of impulse responses by local projections. *American Economic Review*, *95*(1), 161–182.

Kane, T. J. (2003). A quasi-experimental estimate of the impact of financial aid on college-going. NBER Working Papers 9703, National Bureau of Economic Research, Inc.

Kawai, K., & Nakabayashi, J. (2014). Detecting large-scale collusion in procurement auctions. SSRN Electronic Journal.

Keele, L. J., & Titiunik, R. (2015). Geographic boundaries as regression discontinuities. *Political Analysis*, *23*(1), 127–155.

Kong, Y. (2021). Sequential auctions with synergy and affiliation across auctions. *Journal of Political Economy*, *129*(1), 148–181.

Kostol, A. R., & Mogstad, M. (2014). How financial incentives induce disability insurance recipients to return to work. *American Economic Review*, *104*(2), 624–655.

Kuersteiner, G. M., Phillips, D. C., & Villamizar-Villegas, M. (2016). Supplementary material for "the effects of foreign exchange intervention: Evidence from a rule-based policy in Colombia". Borradores de Economia 965, Banco de la Republica de Colombia.

Kuersteiner, G. M., Phillips, D. C., & Villamizar-Villegas, M. (2018). Effective sterilized foreign exchange intervention? Evidence from a rule-based policy. *Journal of International Economics*, *113*, 118–138.

Kumar, A. (2018). Do restrictions on home equity extraction contribute to lower mortgage defaults? Evidence from a policy discontinuity at the texas border. *American Economic Journal: Economic Policy*, *10*(1), 268–97.

Lalive, R. (2008). How do extended benefits affect unemployment duration? A regression discontinuity approach. *Journal of econometrics*, *142*(2), 785–806.

Lee, D. S. (2001). The electoral advantage to incumbency and voters' valuation of politicians' experience: A regression discontinuity analysis of elections to the U.S. NBER Working Papers 8441, National Bureau of Economic Research, Inc.

Lee, D. S. (2008). Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics*, *142*(2), 675–697.

Lee, D. S., & Card, D. (2008). Regression discontinuity inference with specification error. *Journal of Econometrics*, *142*(2), 655–674.

Lee, D. S., & Lemieux, T. (2010). Regression discontinuity designs in economics. *Journal of Economic Literature*, *48*(2), 281–355.

Lee, D. S., & Mas, A. (2012). Long-run impacts of unions on firms: New evidence from financial markets, 1961-1999. *The Quarterly Journal of Economics*, *127*(1), 333–378.

Lee, D. S., & McCrary, J. (2005). Crime, punishment, and myopia. NBER Working Papers 11491, National Bureau of Economic Research, Inc.

Lee, D. S., Moretti, E., & Butler, M. J. (2004). Do voters affect or elect policies? Evidence from the us house. *The Quarterly Journal of Economics*, *119*(3), 807–859.

Lemieux, T., & Milligan, K. (2008). Incentive effects of social assistance: A regression discontinuity approach. *Journal of Econometrics*, *142*(2), 807–828.

Leuven, E., Lindahl, M., Oosterbeek, H., & Webbink, D. (2007). The effect of extra funding for disadvantaged pupils on achievement. *The Review of Economics and Statistics*, *89*(4), 721–736.

Leuven, E., & Oosterbeek, H. (2004). Evaluating the effect of tax deductions on training. *Journal of Labor Economics*, *22*(2), 461–488.

Linden, A., & Adams, J. L. (2012). Combining the regression discontinuity design and propensity score-based weighting to improve causal inference in program evaluation. *Journal of Evaluation in Clinical Practice*, *18*(2), 317–325.

Ludwig, J., & Miller, D. (2007). Does head start improve children's life chances? evidence from a regression discontinuity design. *The Quarterly Journal of Economics*, *122*(1), 159–208.

Luechinger, S., & Roth, F. (2016). Effects of a mileage tax for trucks. *Journal of Urban Economics*, *92*(C), 1–15.

Maddala, G., & Lee, L.-F. (1976). Recursive models with qualitative endogenous variables. *In Annals of Economic and Social Measurement, NBER*, *5*(4), 525–545.

Mariano, B., & Giné, J. A. T. (2015). Creditor intervention, investment, and growth opportunities. *Journal of Financial Services Research*, *47*(2), 203–228.

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

McEwan, P. J., & Shapiro, J. S. (2008). The benefits of delayed primary school enrollment discontinuity estimates using exact birth dates. *Journal of human Resources*, *43*(1), 1–29.

Meyersson, E. (2014). Islamic rule and the empowerment of the poor and pious. *Econometrica*, *82*(1), 229–269.

Mian, A., Sufi, A., & Trebbi, F. (2015). Foreclosures, house prices, and the real economy. *The Journal of Finance*, *70*(6), 2587–2634.

Moscoe, E., Bor, J., & Bärnighausen, T. (2015). Regression discontinuity designs are underutilized in medicine, epidemiology, and public health: A review of current and best practice. *Journal of Clinical Epidemiology*, *68*(2), 132–143.

Müller, T., & Shaikh, M. (2018). Your retirement and my health behavior: Evidence on retirement externalities from a fuzzy regression discontinuity design. Technical report.

Nielsen, H. S., Sørensen, T., & Taber, C. (2010). Estimating the effect of student aid on college enrollment: Evidence from a government grant policy reform. *American Economic Journal: Economic Policy*, *2*(2), 185–215.

Önder, Y. K., & Shamsuddin, M. (2019). Heterogeneous treatment under regression discontinuity design: Application to female high school enrolment. *Oxford Bulletin of Economics and Statistics*, *81*(4), 744–767.

Otsu, T., Xu, K.-L., & Matsushita, Y. (2013). Estimation and inference of discontinuity in density. *Journal of Business & Economic Statistics*, *31*(4), 507–524.

Park, A., Shi, X., Hsieh, C.-t., & An, X. (2015). Magnet high schools and academic performance in China: A regression discontinuity design. *Journal of Comparative Economics*, *43*(4), 825–843.

Pei, Z., & Shen, Y. (2017). *The devil is in the tails: Regression discontinuity design with measurement error in the assignment variable*. Emerald Publishing Limited.

Pence, K. M. (2006). Foreclosing on opportunity: State laws and mortgage credit. *Review of Economics and Statistics*, *88*(1), 177–182.

Perez-Reyna, D., & Villamizar-Villegas, M. (2019). Exchange rate effects of financial regulations. *Journal of International Money and Finance*, *96*(C), 228–245.

Pettersson-Lidbom, P. (2008). Do parties matter for economic outcomes? A regression-discontinuity approach. *Journal of the European Economic Association*, *6*(5), 1037–1056.

Pettersson-Lidbom, P. (2012). Does the size of the legislature affect the size of government? Evidence from two natural experiments. *Journal of Public Economics*, *96*(3), 269–278.

Pitt, M. M., & Khandker, S. R. (1998). The impact of group-based credit programs on poor households in bangladesh: Does the gender of participants matter? *Journal of political economy*, *106*(5), 958–996.

Pitt, M. M., Khandker, S. R., McKernan, S.-M., & Latif, M. A. (1999). Credit programs for the poor and reproductive behavior in low-income countries: Are the reported causal relationships the result of heterogeneity bias? *Demography*, *36*(1), 1–21.

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

Porter, J. (2003). Estimation in the regression discontinuity model. *Unpublished Manuscript, Department of Economics, University of Wisconsin at Madison*, *2003*, 5–19.

Porter, J., & Yu, P. (2015). Regression discontinuity designs with unknown discontinuity points: Testing and estimation. *Journal of Econometrics*, *189*(1), 132–147.

Roberts, M. R., & Sufi, A. (2009). Control rights and capital structure: An empirical investigation. *The Journal of Finance*, *64*(4), 1657–1695.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.

Samarakoon, S., & Parinduri, R. A. (2015). Does education empower women? Evidence from indonesia. *World Development*, *66*, 428–442.

Spenkuch, J. L., & Toniatti, D. (2018). Political advertising and election results. *The Quarterly Journal of Economics*, *133*(4), 1981–2036.

Spiegelman, C. (1976). Two methods of analyzing a nonrandomized experiment "adaptive" regression and a solution to reiersol's problem. *Unpublished dissertation, Northwestern University, Evanston, IL*.

Spiegelman, C. (1977). Fsu statistics report m435.

Stancanelli, E., & Van Soest, A. (2012). Retirement and home production: A regression discontinuity approach. *American Economic Review*, *102*(3), 600–605.

Sween, J., & Campbell, D. T. (1965). A study of the effect of proximally autocorrelated erroron tests of significance for the interrupted time series quasi-experimental design. Northwestern University Department of Psychology.

Thistlethwaite, D. L., & Campbell, D. T. (1960). Regression-discontinuity analysis: An alternative to the ex post facto experiment. *Journal of Educational Psychology*, *51*(6), 309–317.

Trochim, W., & Spiegelman, C. (1980). The relative assignment variable approach to selection bias in pretest-posttest group designs. *Proceedings of the Survey Research Section*, 376–80.

Uppal, Y. (2009). The disadvantaged incumbents: estimating incumbency effects in indian state legislatures. *Public choice*, *138*(1-2), 9–27.

Urquiola, M. (2006). Identifying class size effects in developing countries: Evidence from rural schools in Bolivia. *The Review of Economics and Statistics*, *88*(1), 171–177.

Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. *American Economic Review*, *99*(1), 179–215.

Valencia, F. (2020). Historical econometrics: Instrumental variables and regression discontinuity designs. CEPR Discussion Papers 15208, C.E.P.R. Discussion Papers.

Van der Klaauw, W. (2002). Estimating the effect of financial aid offers on college enrollment: A regression-discontinuity approach. *International Economic Review*, *43*(4), 1249–1287.

Van der Klaauw, W. (2008). Breaking the link between poverty and low student achievement: An evaluation of Title I. *Journal of Econometrics*, *142*(2), 731–756.

Vargas-Herrera, H., & Villamizar-Villegas, M. (2019). Effectiveness of FX intervention and the flimsiness of exchange rate expectations. Borradores de Economia 1070, Banco de la Republica de Colombia.

Venkataramani, A. S., Bor, J., & Jena, A. B. (2016). Regression discontinuity designs in healthcare research. BMJ, 352:i1216.

Viard, V. B., & Fu, S. (2015). The effect of beijing's driving restrictions on pollution and economic activity. *Journal of Public Economics*, *125*, 98–115.

Villamizar-Villegas, M., & Önder, Y. K. (2020). Uncovering time-specific heterogeneity in regression discontinuity designs. Borradores de Economia 1141, Banco de la Republica de Colombia.

Wu, J., Wei, X., Zhang, H., & Zhou, X. (2019). Elite schools, magnet classes, and academic performances: Regression-discontinuity evidence from china. *China Economic Review*, *55*, 143–167.

Yanagi, T. (2014). The effect of measurement error in the sharp regression discontinuity design. *KIER Discussion Paper*, *910*.

Ye, J. (2017). Better safe than sorry? Evidence from Lanzhou's driving restriction policy. *China Economic Review*, *45*, 1–21.

Yu, P. (2012). Identification of treatment effects in regression discontinuity designs with measurement error. *Unpublished*.

Zhang, B., Chen, X., & Guo, H. (2018). Does central supervision enhance local environmental enforcement? Quasi-experimental evidence from China. *Journal of Public Economics*, *164*(C), 70–90.

Zhang, S., Zhong, R., & Zhang, J. (2017). School starting age and academic achievement: Evidence from china's junior high schools. *China Economic Review*, *44*(C), 343–354.

Zhang, Y., Salm, M., & van Soest, A. (2018). The effect of retirement on healthcare utilization: Evidence from china. *Journal of health economics*, *62*, 165–177.

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.