

Joint Inference for the Regression Discontinuity Effect and Its External Validity

Yuta Okamoto^b

Graduate School of Economics, Kyoto University

April 14, 2026

Abstract

The external validity of regression discontinuity designs is crucial for informing policy but is rarely examined in applied work. To advance empirical practice, we propose a joint inference procedure for the treatment effect and its local external validity, captured by the treatment effect derivative (TED), within a robust bias correction framework. We further introduce a locally linear treatment effects assumption, which extends the scope of the TED and enables identification and the construction of a uniform confidence band for extrapolated effects. These methods apply to most empirical studies. Empirical illustrations demonstrate their practical usefulness.

Keywords: Extrapolation, regression discontinuity designs, robust bias correction

JEL Classification: C12, C14, C21

^byuta.okamoto@r.hit-u.ac.jp

I greatly benefited from the constructive suggestions of the co-editor, Ismael Mourifé, and three anonymous referees. This research was conducted while I was a visiting research student at the London School of Economics and Political Science. I am grateful for its hospitality. I also thank Shunsuke Imai for helpful discussions and acknowledge financial support from JST SPRING, Grant Number JPMJSP2110.

1 Introduction

In causal analysis, internal validity and external validity are two central concerns. Internal validity is a necessary condition for drawing causal conclusions, but external validity is equally essential for informing policy. Thus, internally valid causal analysis alone is not sufficient, as emphasized in [Duflo et al. \(2007, Chapter 8\)](#) and [Vivaldi \(2020\)](#). Nevertheless, external validity is often overlooked, as [Peters et al. \(2016\)](#) report for randomized controlled trials, suggesting that empirical studies rarely address this issue explicitly.

Regression discontinuity (RD) designs, one of the most popular and credible empirical strategies, face a similar, if not more severe, challenge. In the RD framework, treatment status changes discontinuously at a cutoff point, enabling identification of the treatment effect at the threshold under a mild smoothness assumption ([Hahn et al., 2001](#)). This attractive feature, however, comes at the cost of uncertain external validity. Standard RD designs provide no information about treatment effects away from the cutoff, thereby limiting the generalizability of the RD estimates regardless of their statistical significance. For example, Figure 1 illustrates two scenarios in which the RD effects at the cutoff are identical, yet their policy implications differ. In Figure 1a, the RD effect plausibly extends beyond the cutoff, whereas in Figure 1b it may be questionable whether one can conclude with confidence that the treatment has a positive effect for everyone. Thus, focusing solely on the treatment effect at the cutoff often leaves the policy implications unclear.

Despite this crucial limitation, we find only 26 papers that discuss external validity among 70 empirical RD studies published in leading economic journals between 2020 and 2024 (details are provided in Section 2). Furthermore, only four studies cite recent theoretical proposals for extrapolation. These results may suggest not only that researchers devote limited attention to external validity, but also that extrapolation strategies proposed in theoretical econometrics are not necessarily well aligned with every empirical context, because such proposals typically require stronger assumptions, informative covariates, or design features such as multiple cutoff points ([Angrist and Rokkanen, 2015](#);

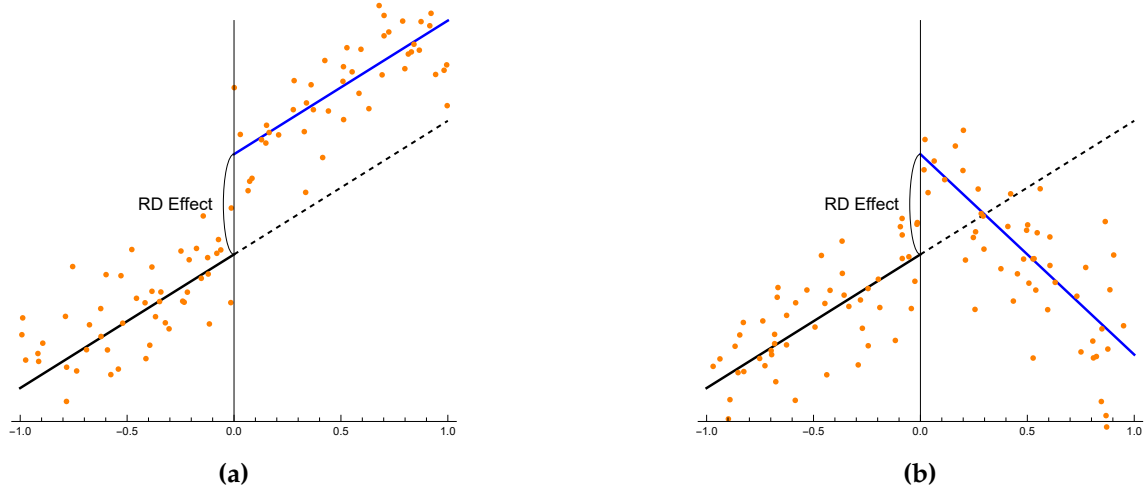


Figure 1: Same RD Effect, Different Policy Implication

Bertanha and Imbens, 2020; Cattaneo et al., 2021; Deaner and Kwon, 2025; Okamoto and Ozaki, 2025). This observation motivates us to develop a simple, practically accessible econometric tool.

To this end, this article develops a simple joint inference procedure by incorporating the local extrapolation method of Dong and Lewbel (2015)—reviewed in the following paragraph—into the widely used robust bias correction (RBC) methods of Calonico et al. (2014). The Dong and Lewbel approach is particularly attractive because it does not require additional covariates or design features and imposes only a weak differentiability assumption; in other words, it can be applied to most empirical settings. Embedding it in the RBC framework ensures compatibility with standard practice: our confidence region is a natural extension of methods already familiar to empirical researchers, remains accessible, and requires no changes to the research design (including bandwidth choice).

To make our proposal concrete, we begin by introducing the Dong and Lewbel strategy. Let $\tau(x)$ denote the treatment effect function with $x = 0$ as the cutoff. Whereas most RD studies focus on identifying and inferring $\tau(0)$, Dong and Lewbel (2015) proposes examining $\tau'(0)$, since this derivative contains information about the local extrapolatability of the RD effect $\tau(0)$. For instance, if $\tau'(0) = 0$, the estimated treatment effect plausibly extends beyond the cutoff, as illustrated in Figure 1a. By contrast, if $\tau'(0) < 0$, the treat-

ment effect may decline at other values of the running variable, raising concerns about the generalizability of the RD effect, as shown in Figure 1b.

In this respect, it is evident that the policy implications of RD studies become clearer when both $\tau(0)$ and $\tau'(0)$ are reported. Unfortunately, however, none of the surveyed studies report the derivative. With the aim of encouraging broader adoption of the [Dong and Lewbel](#) strategy, this article offers two complementary contributions. The first is a joint inference procedure for the pair $(\tau(0), \tau'(0))$, obtained by deriving their joint asymptotic distribution under the RBC framework of [Calonico et al. \(2014\)](#). This RBC-based confidence region should be familiar and accessible to applied researchers and is therefore well-positioned for widespread adoption in empirical work.

Although it is an essential first step, the local external validity indicator $\tau'(0)$ may not provide clear-cut policy implications, particularly when it is non-zero, compared with a more comprehensive extrapolation analysis such as [Cattaneo et al. \(2021\)](#). For instance, what does $\tau'(0) = 0.5$ imply about treatment effects away from the cutoff point? Our second contribution is therefore to introduce a *locally linear treatment effects assumption*, which complements and extends the scope of [Dong and Lewbel](#)'s treatment effect derivative $\tau'(0)$ by enabling more direct extrapolation analysis of RD effects. Under this additional assumption, RD effects away from the cutoff are characterized by the linear function, $\tau(x) = \tau(0) + \tau'(0)x$. This locally linear assumption and resulting identification are *more general* than simply extending the standard RD effect, which effectively restricts $\tau'(0)$ to be zero. Moreover, the limit distribution used to derive the joint confidence region can also be employed to construct a uniform confidence band for the extrapolated RD effects, thereby enabling researchers to conduct a more informative extrapolation analysis that goes beyond merely assessing external validity. Furthermore, we show that the confidence band remains valid for any subregion where the local linearity assumption holds, even if it does not hold over the entire initial interval. That is, even when readers (or policymakers) disagree with researchers about the region over which the assumption is

plausible, they can still perform a valid inference using the reported confidence band for any subinterval. Thanks to this property, our approach does not raise concerns about specification search or ad-hoc tuning, thereby preserving the hallmark transparency of RD designs.

Our proposed approaches require no additional covariates or design features, making them applicable to virtually all RD settings and thereby enhancing the policy relevance of many empirical studies. We illustrate their usefulness with two empirical applications. In the first, we find evidence of external validity and homogeneous treatment effects even away from the cutoff. In the second, local external validity is rejected, suggesting that the standard RD effect should be interpreted with caution; moreover, our extrapolation analysis reveals treatment-effect heterogeneity over the running variable.

Related Literature

RD designs are widely acknowledged as credible quasi-experimental tools, both empirically and theoretically. Empirically, [Brodeur et al. \(2020\)](#) finds that RD papers are less prone to p -hacking and publication bias. [Hyytinen et al. \(2018\)](#) provides empirical evidence showing that the RBC procedure of [Calonico et al. \(2014\)](#) yields inference results that are consistent with experimental estimates at the cutoff. Both works suggest strong internal validity of RD studies. Theoretically, a growing body of research has proposed methods to enhance the internal validity of RD studies. A very short list of recent contributions includes the sharp validity test for fuzzy RD designs by [Arai et al. \(2022\)](#), the manipulation-robust bounding approach of [Gerard et al. \(2020\)](#), the unified falsification test of [Fusejima et al. \(2025\)](#), the analysis of measurement error by [Dong and Kolesár \(2023\)](#), and adjustments for discrete running variables proposed by [Dong \(2015\)](#) and [Kolesár and Rothe \(2018\)](#). See [Cattaneo and Titiunik \(2022\)](#) for a more comprehensive literature review.

In contrast, how carefully external validity is assessed in empirical RD studies re-

mains not fully understood. Theoretically as well, the literature is still developing, as highlighted by [Abadie and Cattaneo \(2018\)](#): “While there is some very recent work that addresses the issue of extrapolation of RD treatment effects [...], more work is certainly needed, covering both methodological and empirical issues.” Recent contributions, other than [Dong and Lewbel \(2015\)](#), include conditional-independence-type extrapolation methods proposed by [Angrist and Rokkanen \(2015\)](#) and [Bertanha and Imbens \(2020\)](#), as well as approaches that exploit additional covariates or multiple cutoff points, such as [Cattaneo et al. \(2016\)](#), [Cattaneo et al. \(2021\)](#), [Deaner and Kwon \(2025\)](#), [Okamoto and Ozaki \(2025\)](#), and [Sun \(2023\)](#). Relatedly, [Bertanha \(2020\)](#) develops a method for estimating average treatment effects under counterfactual treatment-assignment and dose policies in settings with many cutoff points. Furthermore, some studies investigate treatment effect heterogeneity with respect to predetermined covariates ([Hsu and Shen, 2019, 2021](#)). The present article contributes to this second line of literature by providing suggestive survey evidence on recent empirical practice and developing theoretical results for external validity analysis and extrapolation exercises, integrating the seminal yet distinct contributions of [Calonico et al. \(2014\)](#) and [Dong and Lewbel \(2015\)](#). Our locally linear treatment effects assumption has been used in [Dong and Lewbel \(2015\)](#) and [Cerulli et al. \(2017\)](#) without being explicitly stated and is therefore not entirely novel; our contribution is to enhance its practical applicability by providing a uniform inference procedure.

Plan of the Article

Section 2 reviews recent empirical studies to examine the current state of internal and external validity analysis in RD designs. Section 3 presents the main results of the paper, focusing on the sharp RD design. After introducing the setup and notation, Section 3.2 develops a robust local linear confidence region, and Section 3.3 proposes an extrapolation method. Section 4 reports empirical applications to highlight the potential usefulness of our proposed procedure. We also provide simulation studies in Section 5 to investigate

the finite sample properties of our proposed procedures. Finally, Section 6 concludes the paper with practical recommendations. All proofs, along with the generalization to local polynomial confidence regions and the extension to the fuzzy RD design, are provided in the Online Appendix.

2 Empirical Practices in Recent RD Studies

2.1 Sample and Procedure

We begin by investigating how recent RD studies address issues of internal and external validity. We review empirical studies published between 2020 and 2024 in the following leading economics journals: *American Economic Journal: Applied Economics* (AEJ: Applied), *American Economic Journal: Economic Policy* (AEJ: EP), *American Economic Review* (AER), *American Economic Review: Insights* (AERI), *Journal of Human Resources* (JHR), and *Journal of Labor Economics* (JOLE). We identified articles containing the phrase “regression discontinuity” using the EBSCOhost database. After excluding theoretical contributions, we obtain a sample of 70 empirical studies.

To systematically assess whether these studies discuss the external and internal validity of their RD design, and to avoid subjective judgment, we apply the following criteria. External validity is identified by (i) the presence of keywords such as “external validity,” “extrapolat*,” or “transportability” (where the asterisk (*) denotes a wildcard that allows for variations of the root word, e.g., *extrapolate*, *extrapolation*), and (ii) citations to leading contributions including Angrist and Rokkanen (2015), Bertanha and Imbens (2020), Cattaneo et al. (2021), Cerulli et al. (2017), and Dong and Lewbel (2015). Internal validity is assessed using a parallel method. However, discussions of internal validity seldom appear under the explicit phrase “internal validity.” Therefore, in addition to this phrase, we also search for related keywords—such as “continuity,” “manipulat*,” and “placebo”—that suggest a discussion of identification assumptions. We further record citations to

Journal	# of Articles	Internal Validity		External Validity		TED
		Keywords	Citation	Keywords	Citation	
AEJ: Applied	16	16	13	7	1	0
AEJ: EP	16	15	15	4	1	0
AER/AERI	10	8	8	4	0	0
JHR	22	22	15	9	1	0
JOLE	6	6	6	2	1	0
Total	70	67	57	26	4	0

Table 1: Reporting of Internal and External Validity in Empirical RD Studies

Note: The table reports, for each journal, the number of empirical RD studies published between 2020 and 2024 (column 2). Columns 3-4 show the number of articles that mention at least one of the internal validity keywords or cite at least one of the designated references (Calonico et al., 2014; McCrary, 2008). Columns 5-6 report the corresponding counts for external validity, based on the presence of keywords or citations to leading contributions (Angrist and Rokkanen, 2015; Bertanha and Imbens, 2020; Cattaneo et al., 2021; Cerulli et al., 2017; Dong and Lewbel, 2015). The last column reports the number of studies that explicitly present the treatment effect derivative (TED) following Dong and Lewbel (2015).

Calonico et al. (2014) and McCrary (2008), which provide valid inference procedures and a diagnostic test that are essential for internally valid RD analysis. We apply these criteria by using text analysis in R (R Core Team, 2023), which allows us to reproducibly detect whether each article includes the specified keywords or citations.¹

2.2 Results

The survey results are summarized in Table 1. We find that almost all RD studies carefully address the internal validity of their analysis. By contrast, far fewer articles explicitly discuss external validity. In total, only 26 out of 70 articles contain the keywords related to external validity, and only four cite the leading contributions in this area.

This pattern may suggest two possible implications. First, the relatively small number of citations compared to keyword mentions indicates that most existing proposals are not necessarily directly applicable to many empirical RD settings, so that discussions of external validity are often conducted on a case-by-case basis. Second, compared to internal

¹We identify discussions of internal and external validity using fixed keywords and citations, which may under-detect conceptual discussions phrased differently or without the designated citations. Our counts should thus perhaps be interpreted as a conservative lower bound on explicit discussion.

validity, researchers devote relatively little attention to external validity. This is evident from the fact that the procedure of [Dong and Lewbel \(2015\)](#) has not been widely adopted in empirical practice, despite its broad applicability.

This issue is of particular concern. Given the inherently local nature of RD designs, the generalizability of the estimated effect is often unclear. To improve the clarity of their policy implications, it would be desirable to have a simple tool that naturally extends methods routinely used in empirical research. In the next section, we propose such a procedure.

3 External Validity and Extrapolation under Sharp RD

To advance empirical practice, this section proposes a simple inference procedure that builds on [Dong and Lewbel \(2015\)](#). To encourage broader adoption of their strategy, we embed it in the widely used RBC confidence intervals of [Calonico et al. \(2014\)](#). This integration ensures close compatibility with standard practice, making the procedure both accessible to applied researchers and implementable without changes to the research design, including bandwidth choice.

Extending this idea, we also introduce an additional layer of assumptions that enables researchers to conduct a more comprehensive extrapolation analysis, rather than only assessing external validity. The proposed strategy again does not require additional covariates or design features, and is therefore applicable to a wide range of RD studies.

3.1 Setup

We focus on the sharp RD design, while the extension to the fuzzy RD case is provided in the Online Appendix. We have a random sample $(Y_i(0), Y_i(1), X_i) : i = 1, \dots, n$. Here, $Y_i(1)$ denotes the potential outcome under treatment, and $Y_i(0)$ the potential outcome under control. X_i is the running variable with density $f(x)$. Unit i receives the

treatment if $X_i \geq 0$ and is untreated otherwise. Hence, the observed outcome is $Y_i = Y_i(0)\mathbf{1}\{X_i < 0\} + Y_i(1)\mathbf{1}\{X_i \geq 0\}$.

Define the treatment effect function as $\tau(x) := \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x]$. Then, the sharp RD effect is given by $\tau_{\text{SRD}} := \tau(0)$. To assess a local external validity, [Dong and Lewbel \(2015\)](#) introduces the sharp RD treatment effect derivative (TED), which is given by $\tau'_{\text{SRD}} := \tau'(0)$. Under a weak differentiability assumption, these are identified as

$$\begin{aligned}\tau_{\text{SRD}} &= \mu_+ - \mu_-, \quad \mu_+ := \lim_{x \downarrow 0} \mu(x), \quad \mu_- := \lim_{x \uparrow 0} \mu(x), \\ \tau'_{\text{SRD}} &= \mu'_+ - \mu'_-, \quad \mu'_+ := \lim_{x \downarrow 0} \mu'(x), \quad \mu'_- := \lim_{x \uparrow 0} \mu'(x),\end{aligned}$$

where $\mu(x) := \mathbb{E}[Y_i|X_i = x]$ ([Dong and Lewbel, 2015](#)). They also show that both τ_{SRD} and τ'_{SRD} can be estimated from standard local linear regression, but their analysis does not extend to a joint inference procedure for the pair $(\tau_{\text{SRD}}, \tau'_{\text{SRD}})$. Our first goal is to perform valid inference on τ_{SRD} and τ'_{SRD} jointly.

3.2 Robust Local Polynomial Confidence Region

We begin by introducing maintained assumptions, which are fairly general and standard in the literature ([Calonico et al., 2014](#)):

Assumption 1. *In a neighbourhood $(-\kappa_0, \kappa_0)$ around the cutoff:*

- (i) $\mathbb{E}[Y_i^4|X_i = x]$ is bounded, and $f(x)$ is continuous and bounded away from zero.
- (ii) $\mu_+(x) := \mathbb{E}[Y_i(1)|X_i = x]$ and $\mu_-(x) := \mathbb{E}[Y_i(0)|X_i = x]$ are S -times continuously differentiable.
- (iii) $\sigma_+^2(x) := \mathbb{V}[Y_i(1)|X_i = x]$ and $\sigma_-^2(x) := \mathbb{V}[Y_i(0)|X_i = x]$ are continuous and bounded away from zero.

Assumption 2. *The kernel function K is a symmetric, nonnegative, bounded, and continuous second-order kernel with compact support on $[-1, 1]$.*

Following the RD literature (Calonico et al., 2014; Hahn et al., 2001), we employ the local linear regression, while a more general treatment is given in the Online Appendix, Section S2.1. Specifically, we estimate $\hat{\tau}_{\text{SRD}} = \hat{\mu}_+ - \hat{\mu}_-$ and $\hat{\tau}'_{\text{SRD}} = \hat{\mu}'_+ - \hat{\mu}'_-$, where

$$\begin{aligned} (\hat{\mu}_+, \hat{\mu}'_+)^{\top} &= \arg \min_{(b_0, b_1)^{\top} \in \mathbb{R}^2} \sum_{i=1}^n \mathbf{1}\{X_i \geq 0\} (Y_i - b_0 - b_1 X_i)^2 \frac{1}{h} K\left(\frac{X_i}{h}\right), \\ (\hat{\mu}_-, \hat{\mu}'_-)^{\top} &= \arg \min_{(b_0, b_1)^{\top} \in \mathbb{R}^2} \sum_{i=1}^n \mathbf{1}\{X_i < 0\} (Y_i - b_0 - b_1 X_i)^2 \frac{1}{h} K\left(\frac{X_i}{h}\right). \end{aligned}$$

We here note that the *simultaneous* estimation of the mean and first-derivative functions is attractive because the leading biases of $\hat{\mu}_+$ and $\hat{\mu}'_+$ depend on the same quantity $\mu_+^{(2)}$ (e.g., Fan and Gijbels, 1996, Theorem 3.3). Hence, only a single quantity needs to be estimated for bias correction.²

The conditional biases of these local linear estimators are given by

$$\begin{aligned} \mathbb{E} [\hat{\tau}_{\text{SRD}} | X_1, \dots, X_n] - \tau_{\text{SRD}} &= h^2 B_{\text{SRD},0} \left(\mu_+^{(2)} - \mu_-^{(2)} \right) \{1 + o_p(1)\}, \\ \mathbb{E} [\hat{\tau}'_{\text{SRD}} | X_1, \dots, X_n] - \tau'_{\text{SRD}} &= h B_{\text{SRD},1} \left(\mu_+^{(2)} - \mu_-^{(2)} \right) \{1 + o_p(1)\}, \end{aligned}$$

where $B_{\text{SRD},0}$ and $B_{\text{SRD},1}$ are constants depending only on observables. We can estimate the unknown components, $\mu_+^{(2)}$ and $\mu_-^{(2)}$, using the local quadratic estimation with the same kernel K and a possibly different bandwidth b . With these local quadratic estimators, we can estimate the leading bias terms by $h^2 B_{\text{SRD},0} (\hat{\mu}_+^{(2)} - \hat{\mu}_-^{(2)})$ and $h B_{\text{SRD},1} (\hat{\mu}_+^{(2)} - \hat{\mu}_-^{(2)})$, respectively. Then, our bias-corrected estimators for $(\tau_{\text{SRD}}, \tau'_{\text{SRD}})$ are defined as

$$\tilde{\tau}_{\text{SRD}} := \hat{\tau}_{\text{SRD}} - h^2 B_{\text{SRD},0} \left(\hat{\mu}_+^{(2)} - \hat{\mu}_-^{(2)} \right), \text{ and } \tilde{\tau}'_{\text{SRD}} := \hat{\tau}'_{\text{SRD}} - h B_{\text{SRD},1} \left(\hat{\mu}_+^{(2)} - \hat{\mu}_-^{(2)} \right).$$

²Since we also care about the first derivative, a local quadratic fit ($p = 2$) may be appealing under additional smoothness assumptions. A single local quadratic regression then yields both the mean and first-derivative estimators, whose leading biases depend on the third derivative (estimable, e.g., via a local cubic fit). While appealing for asymptotic bias, higher-order polynomials can inflate variance at boundary points, so finite-sample performance need not improve (see, e.g., Gelman and Imbens, 2019, Pei et al., 2022, Ruppert and Wand, 1994, Remark 4). Moreover, bias correction becomes more challenging because it requires estimating higher-order derivatives. We investigate finite-sample properties in Section 5.

The following lemma derives the joint limit of these two estimators:

Lemma 1. *Suppose Assumptions 1-2 hold. If $S \geq 3$, $\max\{h, b\} < \kappa_0$, $n \min\{h^5, b^5\} \times \max\{h^2, b^2\} \rightarrow 0$, $n \min\{h, b\} \rightarrow \infty$, then*

$$\Omega^{-1/2} \text{diag}(1, h) \tilde{\Delta}(\tau_{\text{SRD}}, \tau'_{\text{SRD}}) \rightarrow_d \mathcal{N}\left((0, 0)^\top, \text{diag}(1, 1)\right),$$

provided that Ω is invertible, where

$$\tilde{\Delta}(t, t') := \begin{pmatrix} \tilde{\tau}_{\text{SRD}} - t \\ \tilde{\tau}'_{\text{SRD}} - t' \end{pmatrix}, \quad \Omega := \begin{pmatrix} V_{\text{SRD}} & C_{\text{SRD}} \\ C_{\text{SRD}} & V'_{\text{SRD}} \end{pmatrix},$$

and V_{SRD} , V'_{SRD} , and C_{SRD} are provided in the Online Appendix Section S2.1. Furthermore, Ω is asymptotically invertible. ♣

This is just a straightforward extension of the main theorem of [Calonico et al. \(2014\)](#). Importantly, there is *no* additional unknown components in Ω compared to the asymptotic variance of [Calonico et al. \(2014\)](#). Hence, the same variance estimation procedures proposed in [Calonico et al. \(2014, Section 5\)](#) can be employed in our case. Furthermore, the rate conditions on the bandwidths are the same as those made in [Calonico et al. \(2014\)](#). Hence, we can still use the usual MSE-optimal bandwidth for the treatment effect τ_{SRD} as the main bandwidth h , and that for $\mu_+^{(2)} - \mu_-^{(2)}$ as the bandwidth b for bias estimation, following [Calonico et al. \(2014, Section 4\)](#). That is, an additional bandwidth selection is not needed for valid inference.³

We are now in a position to present the joint inference procedure. Let $\hat{\Omega}$ be a consistent estimator, and define $\hat{\Omega}_h := \text{diag}(1, h^{-1}) \hat{\Omega} \text{diag}(1, h^{-1})$. Due to Lemma 1, we obtain that $\tilde{\Delta}(\tau_{\text{SRD}}, \tau'_{\text{SRD}})^\top \hat{\Omega}_h^{-1} \tilde{\Delta}(\tau_{\text{SRD}}, \tau'_{\text{SRD}}) \rightarrow_d \chi_2^2$. Hence, we can construct an asymptotically valid

³Although the MSE-optimal bandwidth for τ_{SRD} is the option most consistent with standard RD practice, other choices are possible. For example, the asymptotically optimal bandwidth that minimizes $w\text{MSE}[\hat{\tau}_{\text{SRD}}] + (1 - w)\text{MSE}[\hat{\tau}'_{\text{SRD}}]$ for $w \in [0, 1)$ coincides with the MSE-optimal bandwidth for the derivative τ'_{SRD} , which is also valid. See also Section 3.4 for an additional remark and Section 5 for a simulation comparison.

100(1 - α)% confidence region for $(\tau_{\text{SRD}}, \tau'_{\text{SRD}})$ as follows. Let $\alpha \in (0, 1)$ hereafter.

Proposition 1. *Let $\hat{\Omega}$ be invertible and $\hat{\Omega} \rightarrow_p \Omega$. Under the same assumptions in Lemma 1, an asymptotic 1 - α confidence region of $(\tau_{\text{SRD}}, \tau'_{\text{SRD}})$ is given by*

$$\mathcal{R}_{1-\alpha} := \left\{ (t, t')^\top \in \mathbb{R}^2 : \tilde{\Delta}(t, t')^\top \hat{\Omega}_h^{-1} \tilde{\Delta}(t, t') \leq c_{1-\alpha} \right\},$$

where $c_{1-\alpha}$ is the 100(1 - α)-percentile of χ^2_2 distribution. ♣

This confidence region can be represented graphically as an ellipse; see Sections 4 and 5 for illustration. Although somewhat repetitive, it bears emphasis that no additional variance estimation or bandwidth selection is required. Consequently, researchers can incorporate this external validity assessment seamlessly into standard RD analysis without incurring any additional cost.

3.3 Inference under a Locally Linear Treatment Effects Assumption

The confidence region $\mathcal{R}_{1-\alpha}$ is itself a useful object for assessing the external validity of standard RD estimates. However, in many applied policy settings, a more comprehensive extrapolation analysis, such as that of Cattaneo et al. (2021), may be more directly relevant than merely verifying external validity. To this end, we introduce an additional layer of assumptions, which we term the *locally linear treatment effects assumption*.⁴

Assumption 3 (Locally Linear Treatment Effects). *For some $\delta_1, \delta_2 > 0$, the treatment effect function $\tau(x)$ is a linear function of x over the region $[-\delta_1, \delta_2]$.*

Before discussing this assumption, we provide the following identification result of the extrapolated treatment effects:

⁴The idea of imposing additional layer of assumptions is akin to the notion of layered analysis in the partial identification literature (e.g., Manski and Nagin, 1998; Manski, 2011), where progressively stronger yet still acceptable assumptions are imposed to sharpen conclusions. By analogy, in the RD setting, it seems natural to move from the identification at the cutoff under continuity, to local extrapolation under differentiability, and then to further extrapolation under local linearity.

Lemma 2. Suppose that $\mu_+(x)$ and $\mu_-(x)$ are continuously differentiable. Then, under Assumption 3, $\tau(x)$ is identified over $[-\delta_1, \delta_2]$ as $\tau(x) = \tau_{\text{SRD}} + \tau'_{\text{SRD}}x$. ♣

Assumption 3 could be restrictive and may not hold in all empirical settings.⁵ Even so, it provides a natural and tractable starting point for extrapolation analysis, reflecting the common belief that treatment effects are likely to evolve smoothly. Moreover, this extrapolation result is *more general* than simply extending the standard RD effect τ_{SRD} . Hence, it provides a more encompassing perspective than a sole focus on the treatment effect at the cutoff point.

Further, with the help of the additional assumption, Lemma 2 provides a way to intuitively and graphically interpret Dong and Lewbel’s TED. Of course, the TED is a more general quantity and should not be restricted to our interpretation in every scenario. Nevertheless, our framework offers a natural lens through which their TED can be visually understood in practice, and facilitates a more informative extrapolation analysis.

Under the locally linear treatment effects assumption, we can further obtain a useful uncertainty quantification method:

Proposition 2. In addition to the assumptions made in Proposition 1, suppose that Assumption 3 also holds. Then, $\lim_{n \rightarrow \infty} \mathbb{P} [\tau(x) \in \mathcal{U}_{1-\alpha}(x), \forall x \in [-\delta_1, \delta_2]] = 1 - \alpha$, where

$$\mathcal{U}_{1-\alpha}(x) := \left[\tilde{\tau}_{\text{SRD}} + \tilde{\tau}'_{\text{SRD}}x - c_{1-\alpha}^* \sqrt{(1, x) \hat{\Omega}_h(1, x)^\top}, \right. \\ \left. \tilde{\tau}_{\text{SRD}} + \tilde{\tau}'_{\text{SRD}}x + c_{1-\alpha}^* \sqrt{(1, x) \hat{\Omega}_h(1, x)^\top} \right],$$

and $c_{1-\alpha}^*$ is defined by $P(c_{1-\alpha}^*) = 1 - \alpha$, where

$$P(s) := \frac{\ell}{\pi} \left[1 - \exp\left(-\frac{s^2}{2}\right) \right] + \frac{2}{\pi} \int_0^{\pi/2 - \ell/2} 1 - \exp\left(-\frac{s^2}{2 \cos(u)^2}\right) du.$$

⁵Ghosh et al. (2025) recently proposed a similar (and stronger) linear treatment effects assumption in the sharp RD setting (see their Assumption 2). However, they introduced it to improve inference on the standard RD estimand τ_{SRD} and do not consider external validity or extrapolation issues.

Here, $\ell \in [0, \pi]$ denotes the smallest angle between the directions $\hat{v}(-\delta_1)$ and $\hat{v}(\delta_2)$, where $\hat{v}(x) := \hat{\Omega}_h^{1/2}(1, x)^\top / \|\hat{\Omega}_h^{1/2}(1, x)^\top\|$. ♣

Although it may appear obscure at first glance, the confidence band $\mathcal{U}_{1-\alpha}(x)$ is related to the joint confidence region $\mathcal{R}_{1-\alpha}$. Noting that $1 - \exp(-s^2/2)$ is the distribution function of the standard Rayleigh distribution, it follows that $c_{1-\alpha}^* \rightarrow \sqrt{c_{1-\alpha}}$ as $\min\{\delta_1, \delta_2\} \rightarrow \infty$, i.e., $\ell \rightarrow \pi$. In this sense, the confidence band $\mathcal{U}_{1-\alpha}(x)$ admits an alternative interpretation: as the extrapolation interval expands, $\mathcal{U}_{1-\alpha}(x)$ converges to the *projection envelope* $\mathcal{E}_{1-\alpha}(x)$ of the ellipse $\mathcal{R}_{1-\alpha}$ onto the linear functions, where

$$\mathcal{E}_{1-\alpha}(x) := \left[\tilde{\tau}_{\text{SRD}} + \tilde{\tau}'_{\text{SRD}}x - \sqrt{c_{1-\alpha}(1, x)\hat{\Omega}_h(1, x)^\top}, \right. \\ \left. \tilde{\tau}_{\text{SRD}} + \tilde{\tau}'_{\text{SRD}}x + \sqrt{c_{1-\alpha}(1, x)\hat{\Omega}_h(1, x)^\top} \right],$$

which can be regarded as the envelope of the linear extrapolations corresponding to all $(\tilde{\tau}_{\text{SRD}}, \tilde{\tau}'_{\text{SRD}})$ pairs contained in $\mathcal{R}_{1-\alpha}$.⁶

Moreover, we can show that $c_{1-\alpha}^* \rightarrow \Psi^{-1}(1 - \alpha/2)$ as $\max\{\delta_1, \delta_2\} \rightarrow 0$ (and hence $x \rightarrow 0$), where Ψ^{-1} denotes the quantile function of the standard normal distribution. Thus, $\mathcal{U}_{1-\alpha}(x)$ collapses to the usual RBC confidence interval for τ_{SRD} of [Calonico et al. \(2014\)](#). That is, $\mathcal{U}_{1-\alpha}(x)$ can be viewed as a natural extension of the RBC confidence intervals to linear extrapolation analysis.

We conclude this subsection with a remark on the choice of the extrapolation window $[-\delta_1, \delta_2]$. This choice should be guided by the researcher's expertise and research question. At first glance, this may appear to allow room for specification search, but that is

⁶It follows that $c_{1-\alpha}^* \leq \sqrt{c_{1-\alpha}}$, and hence $\mathcal{U}_{1-\alpha}(x) \subseteq \mathcal{E}_{1-\alpha}(x)$. Therefore, $\mathcal{E}_{1-\alpha}(x)$ is a valid confidence band, albeit generally conservative over a finite interval. Intuitively, the inclusion (rather than equality) reflects that different x may have different least-favorable directions: over a wide extrapolation region, many pairs $(\tau, \tau') \in \mathcal{R}_{1-\alpha}$ can be least favorable for some x , whereas on a restricted interval only a subset of directions—formally captured by $\hat{v}(x)$ —matters, yielding a tighter band than the one obtained by mere projection (see the proofs of Proposition 2 and Lemma 3). In practice, however, the difference between the two bands may be small when the extrapolation region is only moderately wide; in our applications they are nearly indistinguishable (not reported).

not the case. First, because the intercept and slope are pinned down by the estimates at the cutoff, the point estimates of the extrapolated line are invariant to the choice of the extrapolation region. Moreover, as shown below, it can be justified to select the largest interval that is plausible given the researcher's institutional knowledge and to report the confidence band over this largest region.

Suppose an economist assumes local linearity over the interval $\mathcal{I}_{\text{EC}}$ and reports the confidence band $\mathcal{U}_{1-\alpha}^{\text{EC}}(x)$ on this interval. A policymaker (or reader), however, may believe that $\mathcal{I}_{\text{EC}}$ is too wide, and that linearity holds only on a smaller interval $\mathcal{I}_{\text{PM}} \subset \mathcal{I}_{\text{EC}}$. If the policymaker is correct, then $\mathcal{U}_{1-\alpha}^{\text{EC}}(x)$ is not valid over the full $\mathcal{I}_{\text{EC}}$; nevertheless, it still satisfies

$$\lim_{n \rightarrow \infty} \mathbb{P} [\tau(x) \in \mathcal{U}_{1-\alpha}^{\text{EC}}(x), \forall x \in \mathcal{I}_{\text{PM}}] \geq 1 - \alpha,$$

which is a direct consequence of the following lemma:

Lemma 3. *Let $c_{1-\alpha}^*(\delta_1, \delta_2)$ denote the critical value defined in Proposition 2 for the extrapolation window $[-\delta_1, \delta_2]$. If $\delta'_1 \leq \delta_1$ and $\delta'_2 \leq \delta_2$, then $c_{1-\alpha}^*(\delta'_1, \delta'_2) \leq c_{1-\alpha}^*(\delta_1, \delta_2)$ holds. That is, the critical value is nondecreasing under nested extensions of the extrapolation window. ♣*

Therefore, the confidence band $\mathcal{U}_{1-\alpha}^{\text{EC}}(x)$ reported in a research article remains valid on any subinterval where Assumption 3 holds, even if the assumption fails on the full interval $\mathcal{I}_{\text{EC}}$. In other words, the policymaker can rely on the reported confidence band to conduct a valid inference, *without access to the underlying research data*. This robustness justifies choosing δ_1 and δ_2 as large as is plausible given the researcher's institutional knowledge, provided that the entire confidence band is transparently reported.

3.4 Remarks

Extrapolation Region. In practice, a researcher must choose (δ_1, δ_2) , which should be guided by the research question and empirical context. That said, depending on the set-

ting, it may be difficult to justify a particular extrapolation window, or determine the “baseline specification” to interpret the result. In such cases, we propose considering $\delta_1 = \delta_2 \approx h$ as a rule of thumb. In finite samples, local linear estimation approximates $\mu_+(x)$ and $\mu_-(x)$ by linear functions over $[0, h]$ and $[-h, 0]$, respectively. It seems therefore natural to adopt a linear treatment effect approximation over the same window, $[-h, h]$, as a benchmark.

Testing the Local Linearity. Assumption 3 has a testable implication $\tau^{(2)}(0) = 0$. This can be tested by the local polynomial regression (see, e.g., [Calonico et al., 2020](#)). In cases where additional information such as multiple cutoffs is available, this assumption can be tested under alternative identification assumptions. We do such exercises in our empirical applications.

MISE-Optimal Bandwidth. Although the MSE-optimal bandwidth for τ_{SRD} is the choice most consistent with standard RD practice, one may instead be interested in the MISE-optimal bandwidth in the context of extrapolation. Note that

$$\begin{aligned} \text{MISE} &= \int_{-\delta_1}^{\delta_2} \mathbb{E} \left[\left\{ (\hat{\tau}_{\text{SRD}} + \hat{\tau}'_{\text{SRD}}x) - (\tau_{\text{SRD}} + \tau'_{\text{SRD}}x) \right\}^2 \right] dx \\ &= \text{MSE} [\hat{\tau}_{\text{SRD}}] \int_{-\delta_1}^{\delta_2} dx + \text{MSE} [\hat{\tau}'_{\text{SRD}}] \int_{-\delta_1}^{\delta_2} x^2 dx + 2\mathbb{E} [(\hat{\tau}_{\text{SRD}} - \tau_{\text{SRD}}) (\hat{\tau}'_{\text{SRD}} - \tau'_{\text{SRD}})] \int_{-\delta_1}^{\delta_2} x dx. \end{aligned}$$

By a standard Taylor-expansion argument, the second term is the dominant term as long as $\max\{\delta_1, \delta_2\} < \infty$. Hence, the asymptotically MISE-optimal bandwidth coincides with the MSE-optimal bandwidth for τ'_{SRD} . Alternatively, one can numerically choose the bandwidth that minimizes the sum of asymptotic approximations to the three terms above. See Section 5 for a simulation comparison of these bandwidth selectors.

Fuzzy RD Design. It is straightforward to extend our framework to the fuzzy RD design. In the fuzzy case, the target parameter is given by $\tau_{\text{FRD}} := \mathbb{E}[Y_i(1) - Y_i(0)|X_i =$

$0, T_i(1) > T_i(0)]$, where $T(0) \in \{0, 1\}$ and $T(1) \in \{0, 1\}$ denote the potential treatment indicator. We observe $T_i = T_i(1)\mathbf{1}\{X_i \geq 0\} + T_i(0)\mathbf{1}\{X_i < 0\}$. The local external validity is then assessed by the fuzzy TED, τ'_{FRD} . Under Assumption A3 of [Dong and Lewbel \(2015\)](#), these are identified as

$$\tau_{\text{FRD}} = \frac{\tau_Y}{\tau_T} := \frac{\mu_{Y,+} - \mu_{Y,-}}{\mu_{T,+} - \mu_{T,-}}, \quad \tau'_{\text{FRD}} = \frac{\tau'_Y \tau_T - \tau_Y \tau'_T}{\tau_T^2},$$

where $\mu_{A,+} := \lim_{x \downarrow 0} \mu_A(x)$, $\mu_{A,-} := \lim_{x \uparrow 0} \mu_A(x)$, and $\mu_A(x) := \mathbb{E}[A_i | X_i = x]$, where A equals either Y or T ([Dong and Lewbel, 2015](#), Theorem 2). The Online Appendix derives the joint asymptotic distribution of $(\tau_{\text{FRD}}, \tau'_{\text{FRD}})$, and hence one can draw a similar joint confidence region. For extrapolation exercises, we may assume the local linearity of $\mathbb{E}[Y_i(1) - Y_i(0) | X_i = x, T_i(1) > T_i(0)]$ in $x \in [-\delta_1, \delta_2]$. Then we can obtain analogous results to those in Lemma 2 and Proposition 2.

Implementation Details. The Online Appendix provides step-by-step implementation details along with an R snippet.

4 Empirical Illustrations

We illustrate the potential usefulness of our proposal through two empirical applications.

4.1 Londoño-Vélez, Rodríguez, and Sánchez (2020)

Empirical Context. [Londoño-Vélez et al. \(2020\)](#) examines the effect of Ser Pilo Paga (SPP), a financial aid program introduced in Colombia in 2014, on students' probability of enrolling in higher education, using a standard RD design.

Among our subsample of interest, consisting of students from metropolitan areas who meet a minimum academic standard ($n = 19,738$), eligibility for the SPP scholarship is determined by a family wealth index. After normalizing the running variable so that the

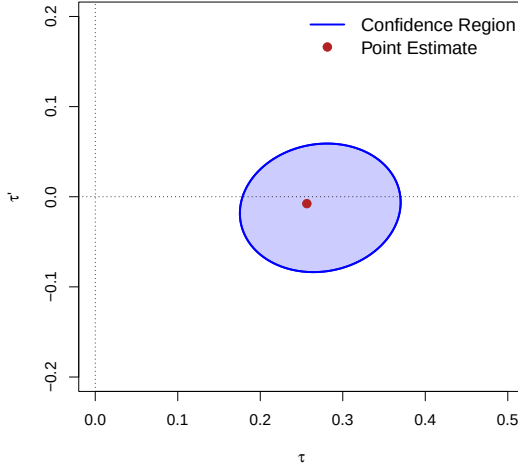
eligibility cutoff equals zero, and multiplying it by -1 to align with the canonical RD convention in which units above the cutoff are treated, the wealth index ranges from -25.94 to 56.86 , with the first quartile at -3.76 and the third quartile at 22.78 . Larger values correspond to poorer households, and students with an index above zero are eligible for the SPP scholarship.

External Validity and Extrapolation Analysis. We begin by assessing the standard RD effect and its local external validity, $(\tau_{\text{SRD}}, \tau'_{\text{SRD}})$, using the joint robust confidence region. For bandwidth selection and standard error estimation, we follow the `rdrobust` package (Calonico et al., 2015, 2017) with resulting bandwidths $(h, b) = (3.8, 7.5)$. Figure 2a reports the 95% confidence region $\mathcal{R}_{0.95}$, where the horizontal axis represents the treatment effect and the vertical axis corresponds to the TED. The figure shows that the SPP has a sizable and statistically significant effect on enrollment probability. At the same time, the confidence region along the TED dimension is relatively tight and centred close to zero, indicating that τ'_{SRD} is small and statistically indistinguishable from zero.

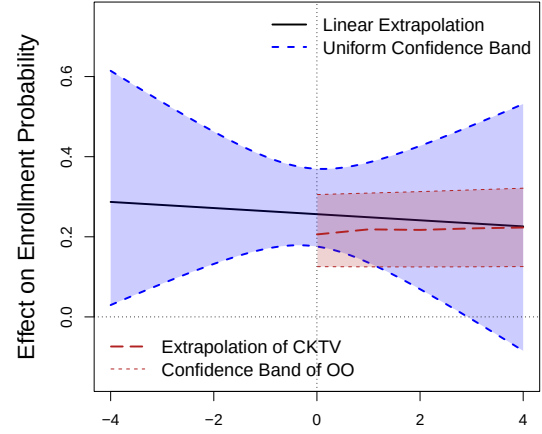
This finding provides direct evidence of the local external validity of the RD effect of the SPP program. Importantly, such an implication could not be drawn from the treatment effect τ_{SRD} alone; it emerges only when the derivative is taken into account. Hence, as emphasized by Dong and Lewbel (2015), examining τ'_{SRD} alongside τ_{SRD} is crucial for clarifying the policy relevance of RD estimates. Our joint inference procedure operationalizes this insight by providing a formal statistical framework that allows both quantities to be assessed simultaneously.

While suggestive, the confidence region $\mathcal{R}_{0.95}$ in Figure 2a provides only local information and is less informative than a full extrapolation analysis. To address this limitation, we impose the locally linear treatment effects assumption over $[-4, 4]$ following the rule of thumb $\delta_1 = \delta_2 \approx h$, with its plausibility briefly discussed at the end of this section.

Figure 2b shows the extrapolated treatment effects along with the uniform confidence



(a) 95% Robust Confidence Region



(b) Linear Extrapolation and Confidence Band

Figure 2: External Validity and Extrapolation of Treatment Effects

Note: **Panel (a).** The dependent variable is a binary indicator of whether the student enrolled in higher education. Hence, the treatment effect reported on the horizontal axis should be interpreted as the effect on the probability of enrollment. **Panel (b).** Linear Extrapolation indicates the point estimates $\hat{\tau}_{\text{SRD}} + \hat{\tau}'_{\text{SRD}}x$, and Uniform Confidence Band shows $\mathcal{U}_{0.95}(x)$. For clarity of presentation, we report only the point estimates of Cattaneo et al. (2021) and the uniform confidence band for the bounds proposed by Okamoto and Ozaki (2025). Larger score values correspond to poorer households.

band $\mathcal{U}_{0.95}(x)$, which delivers clearer implications. Consistent with the robust confidence region in Figure 2a, the uniform confidence band indicates positive treatment effects over $[-4, 3]$, although the effectiveness of SPP at more distant scores remains uncertain. Furthermore, the extrapolated line is rather flat, suggesting the external validity of the local estimate $\tau(0)$.

The positive, seemingly constant effect over $[-4, 3]$ can be economically meaningful. In fact, the extrapolated region $[-4, 3]$ lies around the mode of the running variable's distribution and accounts for more than 16% of our sample. From this perspective, the RD effect is no longer purely local; rather, it is fairly representative—at least for the subpopulation near the mode of the distribution.

To further illustrate the policy implications, consider a simple counterfactual in which the policymaker adjusts the eligibility cutoff for the SPP scholarship. The discussion be-

low assumes that the economy is invariant to the policy change (see [Dong and Lewbel, 2015](#) for details). Suppose the government is considering relaxing the eligibility cutoff to $c^* = -4$. This change would expand the share of students offered the SPP scholarship by approximately 10 percentage points (from the current 66% to about 76% of urban students). Because the estimated treatment effect remains close to the RD effect at the original cutoff even away from the threshold, we project that the resulting increase in higher-education enrollment is approximately $0.10 \times \hat{\tau}_{\text{SRD}} \approx 2.6\%$.

Testing Local Linearity. The key assumption of the previous extrapolation analysis is the local linearity of the treatment effects function $\tau(x)$. As a falsification test, one can test $\tau^{(2)}(0) = 0$. We employ the local cubic smoother and the second derivative is estimated as $\hat{\tau}^{(2)}(0) = 0.009$ with p -value of 0.841 ([Calonico et al., 2020](#)).

In addition, the linearity is also consistent with the findings from alternative approaches that exploit multiple cutoff points, such as [Cattaneo et al. \(2021\)](#) and [Okamoto and Ozaki \(2025\)](#), represented by the red dashed lines and shaded region in Figure 2b.⁷ The linearly extrapolated line on the right-hand side of the cutoff lies within the confidence band of [Okamoto and Ozaki \(2025\)](#) and closely resembles the extrapolated curve of [Cattaneo et al. \(2021\)](#). This concordance supports the plausibility of the locally linear treatment effects assumption in this application.

4.2 Holden (2016)

Empirical Context. [Holden \(2016\)](#) examines the effect of a one-time textbook funding intervention stemming from the 2004 Williams settlement in California. Elemen-

⁷In this application, there exists another group of students from rural regions who face a higher (i.e., more strict) cutoff point. As a result, the regression function of the enrollment probability on the wealth index under the control status can be observed for this additional group even on the right-hand side of the cutoff. [Cattaneo et al.’s \(2021\)](#) strategy achieves point-identified extrapolation by assuming that the regression functions of rural and metropolitan students are “parallel” over $[0, 4]$. By contrast, [Okamoto and Ozaki \(2025\)](#) assumes that the regression function for rural students is monotonically decreasing as the household becomes poorer, and that it lies below that of metropolitan students, thereby yielding partial identification.

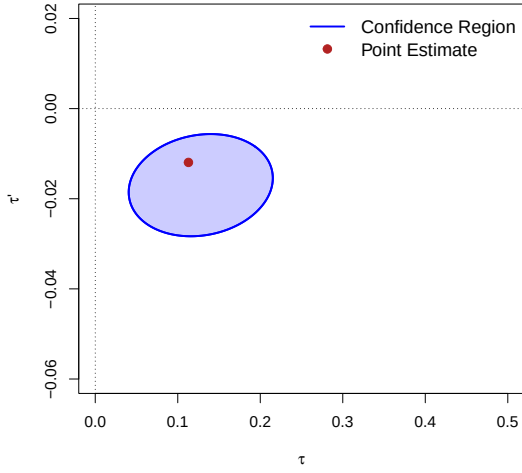
tary schools in the bottom two deciles of the 2003 Academic Performance Index (API)—elementary schools with $\text{API} \leq 643$ —received an extra \$96.90 per student on top of a base \$54.22. Exploiting this discontinuity, the paper implements a sharp RD design to estimate the local causal effect of textbook funding on school-level standardized test performance, finding a positive effect.

We again normalize the API by centering it at the threshold (cutoff = 0) and multiplying by -1 so that values above zero correspond to treatment; in this scale, the running variable ranges from -353 to 268 . Larger values correspond to low performing schools.

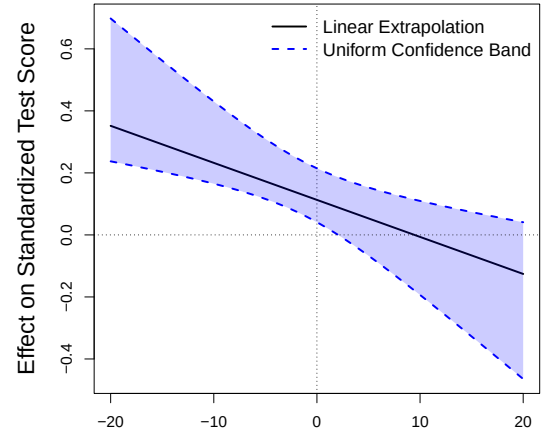
External Validity and Extrapolation Analysis. Figure 3a reports the 95% joint confidence region, using the bandwidths $(h, b) = (20.5, 42.3)$ estimated by `rdrobust` package, to assess the external validity of the RD estimate. Although the estimated treatment effect is positive and statistically significant, we find that the estimated TED is negative and is also statistically different from zero. This may suggest that the treatment effect function is not a constant function even locally, thereby casting a certain concern about the external validity of the standard RD effect.

To more clearly assess the external validity of the RD estimate, we impose a locally linear treatment effects assumption over the window $[-20, 20]$, following the rule of thumb choice $\delta_1 = \delta_2 \approx h$. We compute the implied treatment effects over the region in Figure 3b. The figure yields two salient implications. First, treatment effects can become statistically indistinguishable from zero even for slightly larger score values. Second, effects appear comparatively larger at lower score values within the extrapolation window.

These patterns suggest that the effect identified at the cutoff ($\text{API} = 643$) may not generalize uniformly across the score distribution. In particular, the local RD estimate could overstate the effect for lower-performing schools. One interpretation is that schools struggling with low educational attainment may benefit from more direct and intensive forms of support, such as an after-school program (Muralidharan et al., 2019), and that



(a) 95% Robust Confidence Region



(b) Linear Extrapolation and Confidence Band

Figure 3: External Validity and Extrapolation of Treatment Effects

Note: **Panel (a).** The dependent variable is a normalized test score. **Panel (b).** Linear Extrapolation indicates the point estimates $\hat{\tau}_{\text{SRD}} + \hat{\tau}'_{\text{SRD}}x$, and Uniform Confidence Band shows $\mathcal{U}_{0.95}(x)$. Larger score values correspond to low performing schools.

textbook funding alone may be insufficient to generate sizable improvements, consistent with [Glewwe et al. \(2009\)](#).

Conversely, the extrapolation also hints that better-achieving schools could experience effects that exceed the RD estimate at the cutoff. This pattern would be consistent with complementarities between improved access to instructional materials and pre-existing student ability or school capacity. For instance, higher-performing students may better benefit from additional textbooks ([Glewwe et al., 2009](#)). Likewise, schools with greater instructional capacity (e.g., more experienced teachers) may be better able to integrate new materials into classroom practice, thereby translating improved textbook access into larger test-score gains, consistent with [Mueller \(2013\)](#).

Building on these findings, we can also consider a simple, illustrative policy experiment: if treatment effects are indeed heterogeneous across the running variable, then reallocating a fixed textbook budget could improve cost-effectiveness. In our sample, $\mathbb{P}[-20 < X_i \leq 0] \approx \mathbb{P}[0 \leq X_i < 20] \approx 6\%$, so excluding the interval $[0, 20)$ while cover-

ing $[-20, 0] \cup [20, \infty)$ ⁸ would keep the number of eligible schools roughly comparable to the status quo but shift resources toward score ranges where the extrapolated effects are not attenuated. Under this counterfactual assignment rule, the textbook funding program may yield larger average test-score gains.

Testing Local Linearity. In the absence of the multi-cutoffs as in the previous illustration, we have to rely on the test of $\tau^{(2)}(0) = 0$. We employ the local cubic smoother and the second derivative is estimated as $\hat{\tau}^{(2)}(0) = -0.006$ with p -value of 0.056.

5 Simulation Studies

We present numerical experiments that assess the finite-sample performance of the proposed procedure. We consider the following three data-generating processes (DGPs): $Y_i = \mu_-(X_i)\mathbf{1}\{X_i < 0\} + \mu_+(X_i)\mathbf{1}\{X_i \geq 0\} + \varepsilon_i$, $X_i \stackrel{\text{iid}}{\sim} \text{Uniform}[-5, 5]$, and $\varepsilon_i \stackrel{\text{iid}}{\sim} \mathcal{N}(0, 1)$, where

$$\mu_-(x) = x + \eta(x), \quad \mu_+(x) = 2 + x + \eta(x), \quad (\text{DGP 1})$$

$$\mu_-(x) = x + \eta(x), \quad \mu_+(x) = 2 - x + \eta(x), \quad (\text{DGP 2})$$

$$\mu_-(x) = x - \frac{1}{5}x^2 + \eta(x), \quad \mu_+(x) = 2 - x - \frac{1}{10}x^2 + \eta(x), \quad (\text{DGP 3})$$

and $\eta(x) := \sin(x)/10 + \cos(x)/10$. Plots of the DGPs are provided in the Online Appendix. In every case, the standard RD effect is $\tau_{\text{SRD}} = 2$. In contrast the TED is different: $\tau'_{\text{SRD}} = 0$ for **DGP 1**, and $\tau'_{\text{SRD}} = -2$ for **DGP 2** and **DGP 3**. For the first two DGPs, the treatment effect function is linear in x , while it is not for **DGP 3**.

One-Shot Simulation Illustration. As an illustration, Figure 4 presents a single-shot simulation result of **DGP 2** with a sample size of 1000. Panel (a) shows the confidence re-

⁸Treatment effects over the region $[20, \infty)$ are uncertain in our framework, so we leave the treatment assignment rule for those schools unchanged.

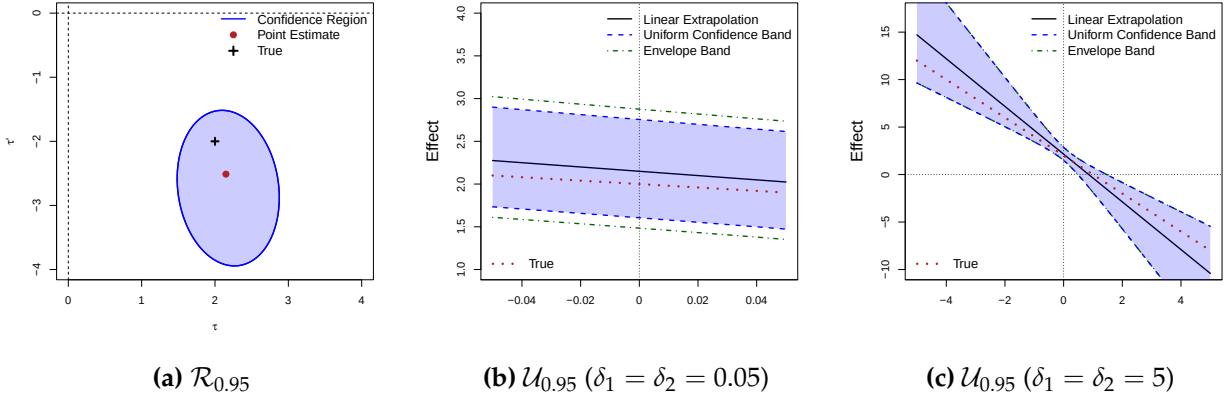


Figure 4: Joint Confidence Region & Confidence Band (DGP 2)

Note: Point Estimate and Linear Extrapolations are reported before bias correction, i.e., $(\hat{\tau}_{\text{SRD}}, \hat{\tau}'_{\text{SRD}})$ and $\hat{\tau}_{\text{SRD}} + \hat{\tau}'_{\text{SRD}}x$.

gion $\mathcal{R}_{0.95}$, which correctly covers the true set $(\tau_{\text{SRD}}, \tau'_{\text{SRD}}) = (2, -2)$. From this picture, we can deduce that the treatment effect at the cutoff is positive, while it decreases as the running variable X increases. Panels (b) and (c) display the uniform confidence band under the locally linear treatment effects assumption over $[-0.05, 0.05]$ and $[-5, 5]$, respectively. In both cases, the uniform confidence band contains the true treatment effect function indicated by the red dotted line. We also observe that the confidence band $\mathcal{U}_{0.95}$ converges to the envelope band $\mathcal{E}_{0.95}$ as (δ_1, δ_2) increase.

Monte Carlo Results. We conduct a Monte Carlo experiment. The main results are reported for $n = 1000$; results for $n = 500$ are deferred to the Online Appendix and lead to the same qualitative conclusions. We consider three symmetric extrapolation regions, $\delta_1 = \delta_2 = \delta \in \{0.05, 1, 5\}$, ranging from a very local window to a wide extrapolation window. Throughout, we consider local linear estimation ($p = 1$) as the baseline, and we additionally report results for local quadratic estimation ($p = 2$) to assess whether using a higher-order polynomial improves finite-sample performance in this setting.

We compare several bandwidth selectors that correspond to the alternative criteria discussed in Section 3.4. First, we use the plug-in MSE-optimal bandwidth for the RD

level effect, denoted by $\hat{h}_{\text{mean}}$, which minimizes the estimated MSE of $\hat{\tau}_{\text{SRD}}$ based on the first-order asymptotic approximation. Second, we use the plug-in MSE-optimal bandwidth for the TED, denoted by $\hat{h}_{\text{deriv}}$, which minimizes the estimated MSE of $\hat{\tau}'_{\text{SRD}}$. Although this is the asymptotically MISE optimal choice for the linear extrapolation, we also consider an alternative MISE-oriented choice, $\hat{h}_{\text{both}}$, obtained by minimizing $\widehat{\text{MISE}} = \widehat{\text{MSE}}[\hat{\tau}_{\text{SRD}}] \int_{-\delta}^{\delta} 1 dx + \widehat{\text{MSE}}[\hat{\tau}'_{\text{SRD}}] \int_{-\delta}^{\delta} x^2 dx$, which may perform better in finite samples. Finally, we report results using the regularized MSE-optimal bandwidths for $\hat{\tau}_{\text{SRD}}$ and $\hat{\tau}'_{\text{SRD}}$, respectively denoted by $\hat{h}_{\text{reg0}}$ and $\hat{h}_{\text{reg1}}$, following [Imbens and Kalyanaraman \(2012\)](#) as implemented in the `rdrobust` package ([Calonico et al., 2017](#)). For bias correction, we use the same MSE-optimal bandwidth for bias estimation as that used by `rdrobust` throughout, regardless of the bandwidth selector used for point estimation.

We first discuss the results for local linear regression ($p = 1$). The left panel of [Table 2](#) reports the empirical coverage probabilities of the proposed confidence region and band, as well as the MISE of the linearly extrapolated effects, based on 1,000 Monte Carlo replications. In all cases, the coverage probability of the joint confidence region $\mathcal{R}_{0.95}$ is excellent and close to the nominal level. The coverage properties are insensitive to the bandwidth choice, reflecting the *robust* nature of the RBC framework. Plus, no single bandwidth selector appears to uniformly dominate the others in terms of empirical coverage probabilities. When the (locally) linear treatment effects assumption holds ([DGP 1](#) and [DGP 2](#)), the uniform confidence band $\mathcal{U}_{0.95}(x)$ also performs well and exhibits a similar coverage accuracy. However, in [DGP 3](#), as this assumption is violated, $\mathcal{U}_{0.95}(x)$ exhibits undercoverage when the extrapolation interval is wide (i.e., when $\delta = 5$). By contrast, when the extrapolation interval is relatively narrow (i.e., $\delta = 1$ and 0.05), the linearity assumption provides a reasonable approximation to the true treatment effects function, resulting in satisfactory coverage.

In terms of MISE, bandwidths that target the MISE criterion ($\hat{h}_{\text{deriv}}$ and $\hat{h}_{\text{both}}$) tend to outperform $\hat{h}_{\text{mean}}$, especially for larger extrapolation windows (i.e., $\delta = 1$ and 5). Never-

theless, these differences are modest when translated into the uniformly averaged point-wise root MSE, $\sqrt{\text{MISE}/(2\delta)}$ (ARMSE), and the RMSE at $x = \pm\delta$ (MSE is reported in the Online Appendix). For $\delta = 1$ under [DGP 1](#) and [DGP 2](#), the ARMSE gap between $\hat{h}_{\text{mean}}$ and $\hat{h}_{\text{both}}$ is only about 0.07. For RMSE at $x = \pm 1$, the gap is 0.09–0.12. Even for the larger window $\delta = 5$, the difference is about 0.29–0.32 in ARMSE and 0.48–0.56 in RMSE, which does not appear substantial relative to the scale of the estimand (see Figure 4c). The regularized bandwidths, $\hat{h}_{\text{reg0}}$ and $\hat{h}_{\text{reg1}}$, improve upon their non-regularized counterparts, perhaps due to the stabilizing effect of regularization. Interestingly, $\hat{h}_{\text{reg0}}$ performs best in terms of finite-sample MISE in our simulations. One contributing factor is that $\hat{h}_{\text{reg0}}$ yields a smaller $\text{MSE}[\hat{\tau}'_{\text{SRD}}]$ than $\hat{h}_{\text{reg1}}$ in our setup. Even so, the resulting differences in ARMSE and RMSE between the two remain limited (e.g., the ARMSE gap is around 0.05 for $\delta = 1$). Overall, both coverage and MISE appear relatively insensitive to the choice of bandwidth selector. That said, given its consistently better MISE performance, we recommend using the usual `rdrobust` bandwidth $\hat{h}_{\text{reg0}}$ as the default choice.

The right panel of Table 2 shows the simulation results for the local quadratic estimation ($p = 2$). Although the main implications remain, the coverage property and MISE both become worse compared to the local linear cases. Hence, except for the case when the sample size is extremely large, the local linear specification would be a better option.

	Local Linear ($p = 1$)					Local Quadratic ($p = 2$)				
	$\hat{h}_{\text{mean}}$	$\hat{h}_{\text{deriv}}$	$\hat{h}_{\text{both}}$	$\hat{h}_{\text{reg0}}$	$\hat{h}_{\text{reg1}}$	$\hat{h}_{\text{mean}}$	$\hat{h}_{\text{deriv}}$	$\hat{h}_{\text{both}}$	$\hat{h}_{\text{reg0}}$	$\hat{h}_{\text{reg1}}$
DGP 1: Coverage										
$\mathcal{R}_{0.95}$	0.943	0.944	0.942	0.936	0.945	0.926	0.928	0.927	0.931	0.934
$\mathcal{U}_{0.95}(\delta = 0.05)$	0.940	0.944	0.941	0.936	0.945	0.928	0.929	0.927	0.924	0.926
$\mathcal{U}_{0.95}(\delta = 1)$	0.942	0.942	0.939	0.935	0.945	0.927	0.928	0.924	0.930	0.935
$\mathcal{U}_{0.95}(\delta = 5)$	0.943	0.944	0.945	0.936	0.945	0.926	0.928	0.927	0.931	0.934
DGP 1: MISE										
$\delta = 0.05$	0.013	0.016	0.013	0.010	0.011	0.047	0.056	0.046	0.017	0.017
$\delta = 1$	0.945	0.868	0.754	0.641	0.756	3.725	3.160	2.623	1.855	2.108
$\delta = 5$	85.93	70.44	67.97	55.59	68.69	353.6	260.1	249.6	191.6	222.6
DGP 2: Coverage										
$\mathcal{R}_{0.95}$	0.944	0.942	0.943	0.940	0.940	0.933	0.938	0.932	0.931	0.929
$\mathcal{U}_{0.95}(\delta = 0.05)$	0.937	0.940	0.937	0.934	0.931	0.927	0.935	0.926	0.930	0.925
$\mathcal{U}_{0.95}(\delta = 1)$	0.944	0.939	0.939	0.939	0.938	0.932	0.936	0.936	0.927	0.927
$\mathcal{U}_{0.95}(\delta = 5)$	0.944	0.942	0.945	0.940	0.940	0.933	0.938	0.937	0.931	0.929
DGP 2: MISE										
$\delta = 0.05$	0.013	0.015	0.013	0.010	0.011	0.037	0.051	0.036	0.016	0.017
$\delta = 1$	0.883	0.815	0.718	0.637	0.745	2.999	2.840	2.222	1.679	1.862
$\delta = 5$	79.75	67.16	64.35	55.49	67.65	287.2	232.5	221.5	170.9	193.1
DGP 3: Coverage										
$\mathcal{R}_{0.95}$	0.937	0.945	0.937	0.942	0.946	0.929	0.925	0.925	0.923	0.927
$\mathcal{U}_{0.95}(\delta = 0.05)$	0.942	0.940	0.942	0.943	0.941	0.925	0.925	0.923	0.928	0.926
$\mathcal{U}_{0.95}(\delta = 1)$	0.939	0.940	0.936	0.938	0.945	0.928	0.924	0.926	0.925	0.928
$\mathcal{U}_{0.95}(\delta = 5)$	0.890	0.884	0.881	0.868	0.904	0.918	0.907	0.907	0.912	0.914
DGP 3: MISE										
$\delta = 0.05$	0.014	0.011	0.014	0.010	0.011	0.045	0.050	0.044	0.017	0.017
$\delta = 1$	0.980	0.742	0.673	0.635	0.764	3.367	3.057	2.430	1.829	2.037
$\delta = 5$	101.5	77.82	75.55	68.33	82.31	326.4	274.9	255.5	201.3	226.7

Table 2: Simulation Results ($n = 1000$)

6 Conclusion and Recommendations

Motivated by our survey suggesting that recent empirical RD studies pay little attention to external validity and that more accessible econometric tools are needed, this article proposes a simple procedure for jointly inferring the RD effect and its local external validity. Building on the insight of [Calonico et al. \(2014\)](#) and [Dong and Lewbel \(2015\)](#), we proposed an RBC-based confidence region for the treatment effect and TED at the cutoff point. We then introduced a locally linear treatment effects assumption, which complements and extends the framework of [Dong and Lewbel \(2015\)](#) by allowing for more direct extrapolation analysis of RD effects. Under this assumption, we further derived a uniform confidence band for the linearly extrapolated effects.

In any RD setting, we recommend reporting both the standard RD effect at the cutoff, τ_{SRD} , and the treatment-effect derivative (TED), τ'_{SRD} , of [Dong and Lewbel \(2015\)](#). Under mild smoothness conditions, the TED serves as a broadly applicable diagnostic for the local external validity of τ_{SRD} , and thus can be reported in virtually all RD studies. Our confidence region $\mathcal{R}_{1-\alpha}$ provides a unified way to make this assessment with formal uncertainty quantification. Moreover, our numerical results support the use of conventional bandwidth choices, so no additional tuning parameters or design modifications are required for valid inference.

When treatment effects away from the cutoff are of substantive or policy interest, the choice of an extrapolation strategy should reflect the available design features and the credibility of any additional assumptions. In the absence of design features—as in our second application—our linear extrapolation approach is appealing because it requires nothing beyond the standard RD setup. Even when multiple cutoffs are available, if the research interest lies to the left of the cutoff, our procedure can still be useful because the multi-cutoff extrapolation methods proposed in the literature are not applicable in that direction ([Cattaneo et al., 2021](#); [Okamoto and Ozaki, 2025](#)). For example, the policy experiment in our first application (relaxing the eligibility threshold) requires extrapo-

lation to the left. By contrast, when the region to the right of the cutoff is of interest, all three approaches are, at least mechanically, applicable. Since each approach relies on certain functional-form assumptions, researchers should choose among them based on the plausibility of the corresponding identifying assumptions. Relatedly, with many cutoffs, detailed counterfactual policy effects (in terms of treatment-assignment rules) can be studied using the methodology developed by Bertanha (2020). Beyond these multi-cutoff approaches, informative covariates can also help extrapolate treatment effects (Angrist and Rokkanen, 2015; Bertanha and Imbens, 2020; Deaner and Kwon, 2025).

Overall, we view joint estimation and inference for $(\tau_{\text{SRD}}, \tau'_{\text{SRD}})$ as a natural default. When extrapolated effects themselves are the target, one can then layer on additional design features or stronger identifying assumptions to draw more informative conclusions.

References

- Abadie, A. and Cattaneo, M. D. (2018). Econometric Methods for Program Evaluation. *Annual Review of Economics*, 10:465–503.
- Angrist, J. D. and Rokkanen, M. (2015). Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away From the Cutoff. *Journal of the American Statistical Association*, 110(512):1331–1344.
- Arai, Y., Hsu, Y., Kitagawa, T., Mourifié, I., and Wan, Y. (2022). Testing identifying assumptions in fuzzy regression discontinuity designs. *Quantitative Economics*, 13(1).
- Bertanha, M. (2020). Regression Discontinuity Design with Many Thresholds. *Journal of Econometrics*, 218(1):216–241.
- Bertanha, M. and Imbens, G. W. (2020). External Validity in Fuzzy Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, 38(3):593–612.

- Brodeur, A., Cook, N., and Heyes, A. (2020). Methods Matter: p-Hacking and Publication Bias in Causal Analysis in Economics. *American Economic Review*, 110(11):3634–60.
- Calonico, S., Cattaneo, M. D., and 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., and Titiunik, R. (2017). rdrobust: Software for Regression-discontinuity Designs. *The Stata Journal*, 17(2):372–404.
- Calonico, S., Cattaneo, M. D., and Titiunik, R. (2014). Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. *Econometrica*, 82(6):2295–2326.
- Calonico, S., Cattaneo, M. D., and Titiunik, R. (2015). rdrobust: An R Package for Robust Nonparametric Inference in Regression-Discontinuity Designs. *The R Journal*, 7:38–51.
- Cattaneo, M. D., Keele, L., Titiunik, R., and Vazquez-Bare, G. (2016). Interpreting Regression Discontinuity Designs with Multiple Cutoffs. *The Journal of Politics*, 78(4):1229–1248.
- Cattaneo, M. D., Keele, L., Titiunik, R., and Vazquez-Bare, G. (2021). Extrapolating Treatment Effects in Multi-Cutoff Regression Discontinuity Designs. *Journal of the American Statistical Association*, 116(536):1941–1952.
- Cattaneo, M. D. and Titiunik, R. (2022). Regression Discontinuity Designs. *Annual Review of Economics*, 14:821–851.
- Cerulli, G., Dong, Y., Lewbel, A., and Poulsen, A. (2017). Testing Stability of Regression Discontinuity Models. In *Regression Discontinuity Designs*, volume 38 of *Advances in Econometrics*, pages 317–339. Emerald Publishing Limited.
- Deaner, B. and Kwon, S. (2025). Extrapolation in Regression Discontinuity Design Using Comonotonicity. <https://arxiv.org/abs/2507.00289>.

- Dong, Y. (2015). Regression Discontinuity Applications with Rounding Errors in the Running Variable. *Journal of Applied Econometrics*, 30(3):422–446.
- Dong, Y. and Kolesár, M. (2023). When can we ignore measurement error in the running variable? *Journal of Applied Econometrics*, 38(5):735–750.
- Dong, Y. and Lewbel, A. (2015). Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models. *The Review of Economics and Statistics*, 97(5):1081–1092.
- Duflo, E., Glennerster, R., and Kremer, M. (2007). Using Randomization in Development Economics Research: A Toolkit. In Schultz, T. P. and Strauss, J. A., editors, *Handbook of Development Economics*, volume 4, chapter 61, pages 3895–3962. Elsevier.
- Fan, J. and Gijbels, I. (1996). *Local Polynomial Modelling and Its Applications*. Chapman & Hall/CRC.
- Fusejima, K., Ishihara, T., and Sawada, M. (2025). A Unified Diagnostic Test for Regression Discontinuity Designs. *Journal of Econometrics*, 251:106074.
- Gelman, A. and Imbens, G. (2019). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, 37(3):447–456.
- Gerard, F., Rokkanen, M., and Rothe, C. (2020). Bounds on Treatment Effects in Regression Discontinuity Designs with a Manipulated Running Variable. *Quantitative Economics*, 11(3):839–870.
- Ghosh, A., Imbens, G., and Wager, S. (2025). PLRD: Partially Linear Regression Discontinuity Inference. <https://arxiv.org/abs/2503.09907>.
- Glewwe, P., Kremer, M., and Moulin, S. (2009). Many Children Left Behind? Textbooks and Test Scores in Kenya. *American Economic Journal: Applied Economics*, 1(1):112–35.

- Hahn, J., Todd, P., and Van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design. *Econometrica*, 69(1):201–209.
- Holden, K. L. (2016). Buy the Book? Evidence on the Effect of Textbook Funding on School-Level Achievement. *American Economic Journal: Applied Economics*, 8(4):100–127.
- Hsu, Y.-C. and Shen, S. (2019). Testing treatment effect heterogeneity in regression discontinuity designs. *Journal of Econometrics*, 208(2):468–486.
- Hsu, Y.-C. and Shen, S. (2021). Testing monotonicity of conditional treatment effects under regression discontinuity designs. *Journal of Applied Econometrics*, 36(3):346–366.
- Hyytinen, A., Meriläinen, J., Saarimaa, T., Toivanen, O., and Tukiainen, J. (2018). When does regression discontinuity design work? evidence from random election outcomes. *Quantitative Economics*, 9(2).
- Imbens, G. and Kalyanaraman, K. (2012). Optimal Bandwidth Choice for the Regression Discontinuity Estimator. *The Review of Economic Studies*, 79(3):933–959.
- Kolesár, M. and Rothe, C. (2018). Inference in Regression Discontinuity Designs with a Discrete Running Variable. *American Economic Review*, 108(8):2277–2304.
- Londoño-Vélez, J., Rodríguez, C., and Sánchez, F. (2020). Upstream and Downstream Impacts of College Merit-Based Financial Aid for Low-Income Students: Ser Pilo Paga in Colombia. *American Economic Journal: Economic Policy*, 12(2):193–227.
- Manski, C. F. (2011). Policy Analysis with Incredible Certitude. *The Economic Journal*, 121(554):F261–F289.
- Manski, C. F. and Nagin, D. S. (1998). Bounding Disagreements About Treatment Effects: A Case Study of Sentencing and Recidivism. *Sociological Methodology*, 28(1):99–137.
- McCrary, J. (2008). Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test. *Journal of Econometrics*, 142(2):698–714.

- Mueller, S. (2013). Teacher experience and the class size effect — experimental evidence. *Journal of Public Economics*, 98:44–52.
- Muralidharan, K., Singh, A., and Ganimian, A. J. (2019). Disrupting Education? Experimental Evidence on Technology-Aided Instruction in India. *American Economic Review*, 109(4):1426–60.
- Okamoto, Y. and Ozaki, Y. (2025). On Extrapolation of Treatment Effects in Multiple-Cutoff Regression Discontinuity Designs. <https://arxiv.org/abs/2412.04265>.
- Pei, Z., Lee, D. S., Card, D., and Weber, A. (2022). Local Polynomial Order in Regression Discontinuity Designs. *Journal of Business & Economic Statistics*, 40(3):1259–1267.
- Peters, J., Langbein, J., and Roberts, G. (2016). Policy evaluation, randomized controlled trials, and external validity—a systematic review. *Economics Letters*, 147:51–54.
- R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Ruppert, D. and Wand, M. P. (1994). Multivariate Locally Weighted Least Squares Regression. *The Annals of Statistics*, 22(3):1346 – 1370.
- Sun, Y. (2023). Extrapolating Away from the Cutoff in Regression Discontinuity Designs. <https://arxiv.org/abs/2311.18136>.
- Vivalt, E. (2020). How Much Can We Generalize From Impact Evaluations? *Journal of the European Economic Association*, 18(6):3045–3089.