

No. 2224

October 2026

A double-differences approach to inference in regression discontinuity designs with a discrete running variable

Georg Graetz
Mattias Nordin

Abstract

We present a new approach to nonparametric inference in regression discontinuity designs where the running variable is discrete and may have few support points. The approach is based on an intuitive and comparatively weak identification assumption: In the absence of treatment, the difference in population mean outcomes at the cutoff is not the largest nor the smallest among differences at any two adjacent support points. Test statistic and confidence interval are constructed from a set of t-statistics based on double differences of the outcome variable, eliminating the need to estimate a conditional expectation function. Simulation evidence suggests that statistical power compares favorably to existing methods. We illustrate our approach by revisiting two studies from the economics of education literature.

Keywords: regression discontinuity design; causal inference; policy evaluation; economics of education
JEL codes: C12; C13; C14; I21

This paper was produced as part of the Labour Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

We gratefully acknowledge financial support from the Institute for Evaluation of Labor Market and Education Policy (IFAU, Dnr 232/2023). We thank Adrian Adermon, Steve Dieterle, Steve Pischke, Fredrik Sävje, Ingeborg Waernbaum as well as seminar and workshop participants at IFAU, University of Bergamo, University of Edinburgh, and University of Glasgow for helpful comments. ChatGPT was used to express test statistics in a more compact form.

Georg Graetz, Uppsala University and Centre for Economic Performance at the London School of Economics. Mattias Nordin, Uppsala University.

Published by
Centre for Economic Performance
London School of Economic and Political Science
Houghton Street
London WC2A 2AE

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means without the prior permission in writing of the publisher nor be issued to the public or circulated in any form other than that in which it is published.

Requests for permission to reproduce any article or part of the Working Paper should be sent to the editor at the above address.

© G. Graetz and M. Nordin, submitted 2026.

1 Introduction

Regression discontinuity (RD) designs have become a standard tool for researchers interested in causal identification. RD designs involve a fixed treatment assignment rule whereby units are treated if and only if they exceed a cutoff value on a specific variable, called the running variable. For instance, students may be offered a place on a prestigious university program if and only if their high school grade point average exceeds a certain value. If the conditional expectation function (CEF) mapping the running variable to the outcome of interest is continuous (absent treatment) around the cutoff, then the causal effect of the treatment on the outcome, for units at the cutoff, is nonparametrically identified. While the estimand is of a local nature, it is often highly policy-relevant.

A key practical challenge that often arises is that the running variable is discrete and takes on only a moderate number of distinct values even as the sample grows large. Examples include students' test scores, rankings, number of employees within a firm, or year of birth.¹ With a discrete running variable, the treatment effect at the cutoff is no longer nonparametrically identified. Even with infinite data, it is impossible to estimate the CEF in an arbitrarily small neighbourhood around the cutoff. To ensure identification and correct inference, it is therefore necessary to place assumptions on the CEF. Existing approaches rely on assumptions that are, arguably, either restrictive or difficult to evaluate (see Section 2).

In this paper, we propose a new approach to inference in RD designs with a discrete running variable, which rests on an intuitive and comparatively weak identification assumption. The motivation for our approach stems from asking what kind of evidence is usually called for, in a discrete RD setting, to conclude that the treatment effect at the cutoff is different from zero: namely, a marked change in the outcome as treatment status changes at the cutoff, relative to changes we see away from the cutoff. This is the notion that we formalize.

Our identification assumption is that in the absence of treatment, the change (first difference) at the cutoff is not the largest, nor the smallest, of all changes in the support of the running variable. To reject the null hypothesis of, for instance, no effect at the cutoff, we require that all *double differences* (the first difference at the cutoff minus the first difference at some other point of the running variable) are statistically significant in the same direction. That is, either all are statistically significantly positive, suggesting a positive treatment effect, or statistically significantly negative, suggesting a negative treatment effect. We prove that this test procedure is conservative in the sense that, if the null hypothesis is true, then the probability of rejection is no higher

¹See Section 2 for recent examples from the literature.

than the chosen significance level.

The test procedure naturally implies a p -value and, by inversion, a confidence interval. Under the identification assumption, the confidence interval is guaranteed not to undercover—it contains the true effect in a fraction of cases at least as large as the specified confidence level.

Although our focus is on statistical inference, the identification assumption implies set-identification of the treatment effect at the cutoff. That is, under the identification assumption, the population CEF implies bounds on the treatment effect. These bounds can be estimated from the range of double differences of the outcome. The confidence interval shrinks towards the true bounds asymptotically.

Our approach can handle multiple cutoffs, non-independent observations, as well as the inclusion of control variables. Its implementation is straightforward, involving estimation of a saturated linear regression model and a series of associated Wald tests.²

As existing methods for optimal bandwidth selection are not valid in our setting, we recommend reporting results for a range of bandwidths. In the standard RD setting with a continuous running variable, narrowing the bandwidth reduces bias but increases variance. The situation is reversed under our approach. A narrower bandwidth implies fewer comparisons that need to be passed to reject the null, which mechanically increases the probability of rejection. Indeed, our identification assumption must be stated with respect to a bandwidth, and is stronger the narrower the bandwidth. A wider bandwidth, on the other hand, weakens the identification assumption but also lowers power.

We present three arguments for the adoption of our approach by practitioners. First, the identification assumption is no stronger, and in some cases strictly weaker, than those of existing methods. Second, implementation is more straightforward, as researchers do not need to take a stand on the smoothness or functional form of the CEF. Third, our approach performs well in terms of statistical power, as we demonstrate in a series of Monte Carlo experiments. In particular, compared with the two approaches for conservative inference in discrete RD settings developed by Kolesár and Rothe (2018), our method is more powerful in scenarios where the CEF is nonlinear, or when there are a limited number of support points but a lot of data at each point.

We illustrate our method by revisiting two studies from the economics of education literature. As an example of settings where reported scores are some step function of underlying raw scores, we consider Graetz et al. (2025) who estimate the effect of SAT scores on college enrollment in Sweden. Their setting features few support points.

²Implementations of our approach for both R and Stata are publicly available at <https://github.com/mattiasnordin/rd2diff>.

Moreover, pre-determined characteristics clearly change between support points, invalidating a local randomization approach (Cattaneo et al. 2024). Neither feature poses a problem for our approach, however. The second example we consider is Oreopoulos (2006) who estimates the effect of an increase in the minimum school-leaving age on earnings in the UK. Our results show that the precision of previous estimates is thanks to imposing parametric assumptions on the CEF. Our approach instead results in wider confidence intervals—though comparable to those obtained by Kolesár and Rothe (2018)—and a lack of evidence that the reform raised earnings. We also show that statistical power increases once we adjust for the gradual phase-in of the reform that is a feature of yearly data.

The paper proceeds as follows. Section 2 describes the challenge posed by discrete RD designs and discusses existing solutions. Section 3 presents our approach. Section 4 discusses results from Monte Carlo experiments. Section 5 revisits two applications from the literature and Section 6 offers a concluding discussion.

2 Problem formulation and existing approaches

A brief look at recent literature confirms that RD designs with a discrete running variable are ubiquitous. Some examples of discrete running variables among recently published articles include test scores (Figlio and Ozek, 2024; Londoño-Vélez et al., 2026; Mountjoy, 2026), month or year of birth (Rabaté et al., 2024; Cao et al., 2026), number of employees (Kamal et al., 2026), R&D grant applicant ranking (Howell et al., 2025), and energy efficiency rating (Sejas-Portillo et al., 2025). Most of these articles apply the RDD toolkit developed for continuous running variables to their main specification and address discreteness only in passing, even when the running variable has few support points. Two articles address discreteness by clustering at the level of the running variable, a practice criticized by Kolesár and Rothe (2018).

To appreciate the challenges that a discrete running variable poses for RD designs, recall first how identification is achieved in the standard case where the running variable is continuous. Let this running variable be denoted by X , and let treatment W be assigned as $W = \mathbb{1}\{X \geq c\}$, where $\mathbb{1}$ is the indicator function. The counterfactual conditional expectation functions for treated and untreated outcomes, $\mathbb{E}(Y(1)|X)$ and $\mathbb{E}(Y(0)|X)$ are assumed to be continuous at the cutoff c . The treatment effect at the cutoff, τ_c , can be identified because under continuity,

$$\tau_c = \mathbb{E}(Y(1)|X = c) - \mathbb{E}(Y(0)|X = c) = \lim_{X \downarrow c} \mathbb{E}(Y|X) - \lim_{X \uparrow c} \mathbb{E}(Y|X). \quad (1)$$

The result is illustrated by Figures 1a and 1b. Points A and B correspond to

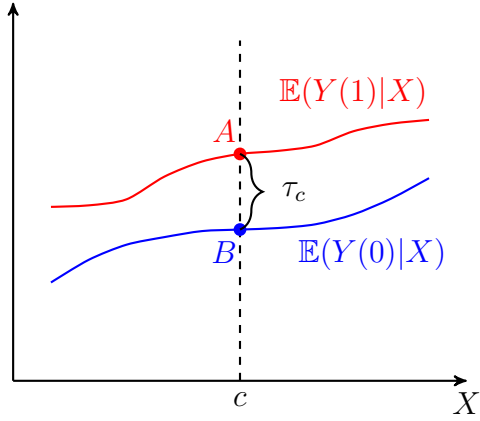
$\mathbb{E}(Y(1)|X = c)$ and $\mathbb{E}(Y(0)|X = c)$ respectively, and the vertical distance between the two points equals τ_c . In terms of $\mathbb{E}(Y|X)$, which is what can be estimated from data, point C is the limit from above, $\lim_{X \downarrow c} \mathbb{E}(Y|X)$, whereas point D is the limit from below, $\lim_{X \uparrow c} \mathbb{E}(Y|X)$. As long as the density of X is non-zero around the cutoff, the treatment effect at the cutoff is nonparametrically identified. Loosely speaking, with infinite data and a continuous running variable, there will be enough observations arbitrarily close to the cutoff to pin down points C and D in Figure 1b.

With a discrete running variable, where the number of support points is fixed, nonparametric identification no longer obtains, as illustrated by Figures 1c and 1d. While point A will be identical to point C , which can be estimated, point B cannot be nonparametrically identified. Instead, it is necessary to make assumptions about the CEF in order to identify τ_c . We discuss four existing approaches to solve this problem.

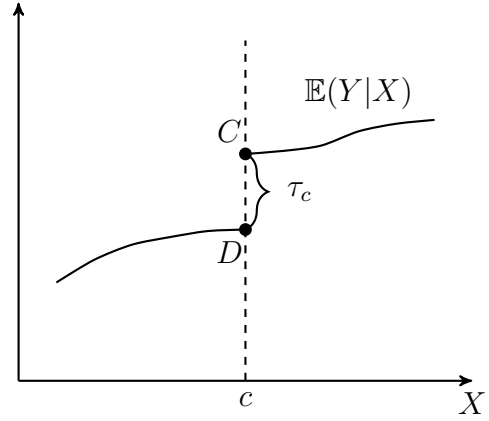
First, researchers may assume that the true CEF, locally within some bandwidth, follows a particular functional form and use that functional form to extrapolate to the cutoff from the left. A natural, but strong, assumption is that of linearity. While linear approximations may be useful in other contexts, here the objective is extrapolation to a point outside the support of the regression function. If the true CEF is nonlinear, this extrapolation will fail. Point estimators will be biased, confidence intervals will under-cover, and hypothesis tests will over-reject. Employing a weighting scheme with a triangular kernel may alleviate, but does not eliminate, this problem. Assuming a more flexible functional form risks over-fitting, especially with few support points.

To guard against misspecification of the CEF, researchers may prefer to state a bound on the deviation of the true CEF from the assumed form. This is one of the approaches suggested by Kolesár and Rothe (2018). They derive honest confidence intervals with conservative coverage properties. To do so, they consider a latent version of the CEF which is twice differentiable, and assume that the second derivative of this latent CEF is bounded by some constant, specific to the particular application. While it is possible to use a data-driven approach of selecting this constant (see Armstrong and Kolesár 2020), such an approach requires many support points and will often be very conservative. Kolesár and Rothe (2018) instead advise the applied researcher to specify the constant a priori based on domain-specific knowledge. A drawback is that it may be difficult to do so in a transparent way, especially in settings where the running variable is measured on an ordinal scale (such as ranks), where the construct of a latent CEF is unappealing.

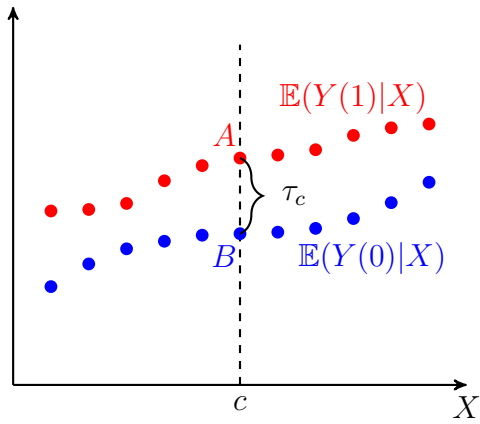
Second, the local randomization approach of Cattaneo et al. (2024) relies on the assumption that the CEF is flat in a neighbourhood around the cutoff (that is, locally constant). The advantage of this approach is that the full toolbox from the analysis of randomized experiments is available. For instance, with this approach one can just



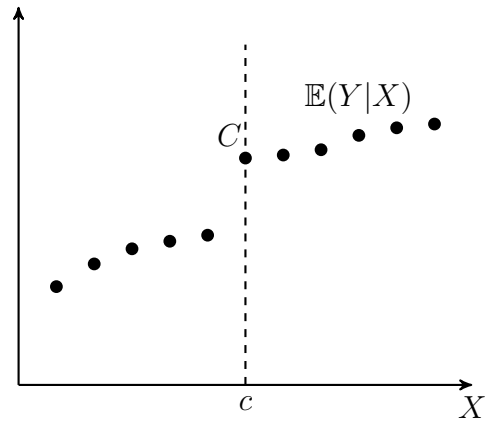
(a) Continuous: Potential outcomes



(b) Continuous: Observed outcomes



(c) Discrete: Potential outcomes



(d) Discrete: Observed outcomes

Figure 1: RD designs with continuous and discrete running variables

compare the mean outcome for units at the cutoff with the mean outcome of units at the support point just left of the cutoff. The drawback is that this assumption is very strong and unlikely to be fulfilled in many settings. We provide evidence against it in one of our applications. A related approach is suggested by Eckles et al. (2025) who study situations where the running variable is a noisy measure of an underlying variable, where the noise can be used in a design-based approach to mimic randomization. For this approach to work, the researcher needs to have knowledge of the noise distribution.

The third approach is to use permutation inference, by “moving around” the cutoff and comparing the estimated treatment effect at the true cutoff to estimates at placebo cutoffs (see, for instance, Section 4.3 in Cattaneo and Titiunik 2022, Ganong and Jäger 2018, and Graetz et al. 2025). Permutation inference is justified if the cutoff location can be viewed as being randomly chosen among a large set of possible locations, an assumption that is often dubious in policy contexts, where cutoff location is chosen for a particular reason. For permutation inference to work, not only is it necessary that what happens at the cutoff, absent treatment, cannot stand out in expectation, but balance is also required in higher moments. For instance, if the outcome is heteroskedastic such that variance of the outcome is higher at the cutoff, then inference will be incorrect. A further drawback of permutation inference is that the resolution of the p -value is constrained by the number of support points.

A fourth approach is to bound the specification error at the cutoff. Kolesár and Rothe (2018) construct confidence intervals under the assumption that the estimated CEF fits no worse at the cutoff than elsewhere within the chosen window. An advantage of this approach is that it is not necessary to make any assumption about a latent CEF outside of the support points. However, we identify two drawbacks with this approach. First, with local linear regression and a uniform kernel (the case Kolesár and Rothe (2018) focus on) the fit at the cutoff tends to be poor. This is because a linear approximation of a nonlinear function fitted with OLS will often fit poorly at the outer support points, which is exactly where the cutoff is located. Second, the assumption is not an assumption about the relationship between the outcome and the running variable (that is, about the CEF), but rather about the researcher’s ability to estimate this relationship. That is, the researcher must employ a priori reasoning to both choose a functional form for the CEF, and to assess how well the CEF thus estimated will fit at the cutoff compared to at other points of the running variable.

Compared to the approaches discussed, we believe an important advantage of our method lies in its identification assumption, which is intuitive, general, and weak. It is intuitive because identification simply requires that the untreated CEF does not exhibit unique behavior at the cutoff. It is general because it does not necessitate

context-dependent parameterization. It is no stronger than the identification assumptions underlying any of the existing approaches, and in several cases strictly weaker. Further advantages are that our method does not require assuming an underlying continuous running variable, and that a low number of support points does not give rise to a low-resolution issue.

A feature that we share with some of the existing approaches is that, while we develop the familiar inference machinery (hypothesis test, p -value, and confidence interval), we do not provide point identification. Instead, we provide bounds on the treatment effect and thus achieve partial identification. Our confidence interval converges asymptotically to the bounds implied by the population CEF.

Our identification assumption is a kind of “bounded-variation assumption” in the terminology of Manski and Pepper (2018), who perform partial identification in a difference-in-differences (DiD) setting, but do not provide tools for statistical inference. Building on their work, Rambachan and Roth (2023) develop two approaches to statistical inference in DiD settings when parallel trends are violated, one using moment inequalities, and one using fixed-length confidence intervals. In contrast, our approach to inference is based on a set of t -statistics.

3 Theory

3.1 Setup and identification assumption

The running variable X has a fixed, finite, number of support points, $G \geq 3$. We denote the possible values X can take as $x_1, x_2, \dots, x_g, \dots, x_{G-1}, x_G$. The running variable may be measured on an ordinal scale, in which case X ranges from 1 to G in order.

W is a treatment indicator equaling one if treated and zero if non-treated. We focus on the sharp RD setting where treatment is deterministically assigned according to a fixed cutoff value x_c , that is, $W = \mathbb{1}\{X \geq x_c\}$, where c is an integer such that $2 \leq c \leq G$. The potential outcome if treated is $Y(1)$ and if non-treated it is $Y(0)$. We assume SUTVA and write the observed outcome variable as $Y = (1-W)Y(0) + WY(1)$.

Data $\{X_i, Y_i, W_i\}_{i=1}^n$ are randomly sampled from some superpopulation, with a fixed sample size, n . As sampling is the sole source of randomness, inference will be made to population parameters, as is standard in RD settings. The probability that unit i has a value of the running variable x_g is denoted $p_g := \Pr(X_i = x_g)$, and we assume that $0 < p_g < 1$ for all g . The number of units for which $X = x_g$ in the sample

is N_g , where $n = \sum_{g=1}^G N_g$.³ By the law of large numbers, $N_g/n \xrightarrow{p} p_g$. The outcome data are collected in a $n \times 1$ vector, $\mathbf{Y}$.

Define $\mathcal{G} := \{2, \dots, G\} \setminus c$ and denote outcome means conditional on support point $g \in \mathcal{G}$ by $\mu(x_g) := \mathbb{E}(Y_i | X_i = x_g)$. Note that $\mu(x_g) = \mathbb{E}(Y_i(0) | X_i = x_g)$ when $g < c$ and $\mu(x_g) = \mathbb{E}(Y_i(1) | X_i = x_g)$ when $g > c$. Furthermore, define $\mu(x_c) := \mathbb{E}(Y_i(1) | X_i = x_c)$ and $\mu_0(x_c) := \mathbb{E}(Y_i(0) | X_i = x_c)$. The target parameter is the average treatment effect at the cutoff,

$$\tau_c := \mathbb{E}(Y_i(1) | X_i = x_c) - \mathbb{E}(Y_i(0) | X_i = x_c) = \mu(x_c) - \mu_0(x_c). \quad (2)$$

For expositional simplicity, we focus on the case where X is equally spaced, but treat the more general case in Section 3.7. Our key assumption is that, if there was no treatment at the cutoff, the change at the cutoff cannot be the smallest nor largest.

Assumption 1. $\min_{g \in \mathcal{G}} \{\mu(x_g) - \mu(x_{g-1})\} \leq \mu_0(x_c) - \mu(x_{c-1}) \leq \max_{g \in \mathcal{G}} \{\mu(x_g) - \mu(x_{g-1})\}$.

The assumption says that there must exist at least one other support point where the change is weakly smaller than the untreated change at the cutoff, as well as at least one other support point where the change is weakly larger than the untreated change at the cutoff.

Under Assumption 1, the only reason for a uniquely small or large increase in population mean outcomes at the cutoff, *relative to the change at other points of the running variable*, is a non-zero treatment effect. If in the data, the change at the cutoff is larger, or smaller, than the changes at all other values of the running variable, then there is evidence of a non-zero treatment effect at the cutoff. This idea forms the basis of our statistical test. Just as with the permutation test approach discussed in the previous section, the intuition behind our approach is that if what happens at the cutoff stands out compared to at other values of the running variable, then that is evidence against the null that there is no treatment effect. Different from the permutation test, however, is that we do not need to assume that what happens at the cutoff is comparable to what happens at every other point of the running variable. Instead, it is sufficient to make the much weaker assumption that the change is not the smallest nor largest without treatment.

Note that the assumption does not only concern what happens absent treatment, since $\mu(x_g)$ also depends on a local treatment effect for $x_g > x_c$. An alternative would be to instead restrict the data to $x_g \leq x_c$, which is feasible as long as $c > 2$. In that case, the assumption does not say anything about treatment effects. We do not

³We could also consider the case where the number of units at each support points is fixed (i.e., stratified sampling). The theoretical results are the same in that case.

impose this restriction for three reasons. First, recall that “untreated” and “treated” in practice do not necessarily imply “not subject to an intervention” and “subject to an intervention”. Instead, they strictly imply only a *change* in the intervention or policy. Second, using data to the right of the cutoff weakens the identification assumption. Finally, once there are multiple cutoffs, the restriction would mean that we can only study the very first cutoff.

It is worth exploring how Assumption 1 relates to statements about changes in absolute value terms. If Assumption 1 holds, then

$$|\mu_0(x_c) - \mu(x_{c-1})| \leq \kappa, \quad \kappa := \max_{g \in \mathcal{G}} \{|\mu(x_g) - \mu(x_{g-1})|\} \quad (3)$$

also holds. In other words, Assumption 1 implies that the untreated change at the cutoff is not the largest in absolute value among all first differences of the CEF. Our identification assumption therefore restricts the smoothness of the CEF in the sense of Lipschitz continuity, as does the approach bounding the second derivative proposed by Kolesár and Rothe (2018). However, with our approach, the restriction can be inferred from data.

Note also that the opposite implication is false. Equation (3) does *not* imply Assumption 1, that is, Assumption 1 is stronger than equation (3). We prefer the stronger assumption as it enables our approach to detect a treatment effect that is of a sign opposite to the general trend in the CEF. For instance, suppose that $\mu(x_g)$ is a monotonically increasing function. If $\mu_0(x_c) - \mu(x_{c-1})$ is positive but smaller than all other $\mu(x_g) - \mu(x_{g-1})$, then equation (3) holds but Assumption 1 is violated. Assumption 1 forbids such a scenario in order for the approach to be capable of detecting treatment effects of a sign opposite to the trend.

While the main focus of this paper is statistical inference, we note that under Assumption 1, the target parameter τ_c is set-identified. To see this, first re-write Assumption 1 as

$$\min_{g \in \mathcal{G}} \{\mu(x_g) - \mu(x_{g-1})\} \leq \mu(x_c) - \mu(x_{c-1}) - \tau_c \leq \max_{g \in \mathcal{G}} \{\mu(x_g) - \mu(x_{g-1})\}. \quad (4)$$

Next, define the double difference $\delta_g := \mu(x_c) - \mu(x_{c-1}) - (\mu(x_g) - \mu(x_{g-1}))$, which compares the change at the cutoff with the change at some other point $g \in \mathcal{G}$, and which features prominently throughout the rest of the paper. By subtracting $\mu(x_c) - \mu(x_{c-1})$ from all sides of the inequality signs in equation (4), we obtain bounds on the

treatment effect at the cutoff,⁴

$$\min_{g \in \mathcal{G}} \{\delta_g\} \leq \tau_c \leq \max_{g \in \mathcal{G}} \{\delta_g\}, \quad (5)$$

which are estimable because each δ_g can be estimated from data (see Section 3.4).

3.2 Hypothesis test

Null and alternative hypotheses are

$$H_0 : \tau_c = \tau^0, \quad H_1 : \tau_c \neq \tau^0, \quad (6)$$

where τ^0 is the hypothesized average treatment effect at the cutoff under the null (which in most applications will be zero). Under this null, we therefore have that $\mu(x_c) = \mu_0(x_c) + \tau^0$. Plugging this expression into Assumption 1 and following similar steps as in the derivation of equation (5), we obtain

$$\min_{g \in \mathcal{G}} \{\delta_g - \tau^0\} \leq 0 \leq \max_{g \in \mathcal{G}} \{\delta_g - \tau^0\}. \quad (7)$$

Suppose we have an estimator of δ_g , D_g , such that $D_g \stackrel{a}{\sim} \mathcal{N}(\delta_g, \mathbb{V}(D_g))$, and that S_g^2 is a consistent estimator of $\mathbb{V}(D_g)$. To test whether the change at the cutoff differs from the change at some other point g , a natural test statistic is

$$T_g(\mathbf{y}) := \frac{D_g - \tau^0}{S_g}, \quad (8)$$

where $\mathbf{y}$ is the realization of $\mathbf{Y}$ for a given sample of data. Consider performing a series of one-sided tests of the null $\delta_g \leq \tau^0$, one for each $g \in \mathcal{G}$. If all of these tests are rejected, then that is evidence that $\tau_c > \tau^0$. Similarly, if all one-sided tests of the null $\delta_g \geq \tau^0$ are rejected, then that is evidence that $\tau_c < \tau^0$. We can combine all those tests to test the null that $\tau_c = \tau^0$.

We define the extrema of the statistics,

$$T_{\min}(\mathbf{y}) := \min_{g \in \mathcal{G}} \{T_g(\mathbf{y})\}, \quad T_{\max}(\mathbf{y}) := \max_{g \in \mathcal{G}} \{T_g(\mathbf{y})\}. \quad (9)$$

The statistic which forms the basis for our hypothesis test is

$$\mathcal{T}(\mathbf{y}) := \min \{|t| : T_{\min}(\mathbf{y}) \leq t \leq T_{\max}(\mathbf{y})\}. \quad (10)$$

⁴The derivation uses that $\min_{g \in \mathcal{G}} \{-\delta_g\} = -\max_{g \in \mathcal{G}} \{\delta_g\}$ and $\max_{g \in \mathcal{G}} \{-\delta_g\} = -\min_{g \in \mathcal{G}} \{\delta_g\}$.

This definition says that $\mathcal{T}(\mathbf{y}) = 0$ if not all $T_g(\mathbf{y})$ have the same sign. Otherwise, it is the smallest absolute value of $T_g(\mathbf{y})$. Let $\alpha \in [0, 1]$ denote the significance level of our test and define the critical value $z_{\alpha/2} := \Phi^{-1}(1 - \alpha/2)$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal CDF. We can now state our main result.

Proposition 1. *A test of the null that $\tau_c = \tau^0$ against the alternative $\tau_c \neq \tau^0$ with decision rule of rejecting the null if and only if $\mathcal{T}(\mathbf{y}) > z_{\alpha/2}$, is a level- α test. That is, under the null,*

$$\Pr(\mathcal{T}(\mathbf{Y}) > z_{\alpha/2}) \leq \alpha. \quad (11)$$

The p -value associated with the test is $p(\mathbf{y}) = 2(1 - \Phi(\mathcal{T}(\mathbf{y})))$.

Proof. We first prove that, under the null, $\Pr(T_{\min}(\mathbf{Y}) > z_{\alpha/2}) \leq \alpha/2$. Take some $g \in \mathcal{G}$ such that $\tau^0 - \delta_g \geq 0$, existence of which is guaranteed by Assumption 1. Let $Z_g := (D_g - \delta_g)/S_g$ and note that $Z_g \stackrel{a}{\sim} \mathcal{N}(0, 1)$ by the regularity conditions we imposed on D_g and S_g . Hence,

$$\Pr(T_{\min}(\mathbf{Y}) > z_{\alpha/2}) \leq \Pr\left(\frac{D_g - \tau^0}{S_g} > z_{\alpha/2}\right) = \Pr\left(Z_g > z_{\alpha/2} + \frac{\tau^0 - \delta_g}{S_g}\right) \leq \frac{\alpha}{2}, \quad (12)$$

where the first inequality is by definition of $T_{\min}(\mathbf{y})$. Assumption 1 also says that there must exist a $g' \in \mathcal{G}$ such that $\tau^0 - \delta_{g'} \leq 0$, and so by analogous reasoning we obtain $\Pr(T_{\max}(\mathbf{Y}) < -z_{\alpha/2}) \leq \alpha/2$. Finally, because $T_{\max}(\mathbf{y}) \geq T_{\min}(\mathbf{y})$, we have

$$\Pr(\mathcal{T}(\mathbf{Y}) > z_{\alpha/2}) = \Pr(T_{\min}(\mathbf{Y}) > z_{\alpha/2}) + \Pr(T_{\max}(\mathbf{Y}) < -z_{\alpha/2}) \leq \alpha, \quad (13)$$

which establishes that the test has level α .⁵

A valid p -value has the property that $\Pr(p(\mathbf{Y}) \leq \alpha) \leq \alpha$ for every $\alpha \in [0, 1]$. By simple manipulations we obtain

$$\Pr(p(\mathbf{Y}) \leq \alpha) = \Pr(2(1 - \Phi(\mathcal{T}(\mathbf{Y}))) \leq \alpha) = \Pr(\Phi(\mathcal{T}(\mathbf{Y})) \geq \Phi(z_{\alpha/2})) \leq \alpha, \quad (14)$$

where the last inequality follows from equation (13) since $\Phi(\cdot)$ is a monotone function. This establishes that $2(1 - \Phi(\mathcal{T}(\mathbf{y})))$ is a valid p -value. $\square$

An alternative way of proving the proposition would be to first use the union-intersection method, where a series of one-sided tests are formed to test whether $\delta_g > \tau^0$, and another series of tests are formed to test whether $\delta_g < \tau^0$. Then, in a

⁵A level- α test is a test which rejects the null with at most probability α under the null. This definition is distinct from a size- α test which rejects the null with probability exactly α .

second step, the intersection-union method can be used to combine these two sets of tests.

3.3 Confidence interval

To obtain a confidence interval, we invert the test.

Lemma 1. *The interval*

$$C(\mathbf{y}) = \left[\min_{g \in \mathcal{G}} \{D_g - z_{\alpha/2} S_g\}, \max_{g \in \mathcal{G}} \{D_g + z_{\alpha/2} S_g\} \right], \quad (15)$$

is a confidence interval with confidence coefficient $1 - \alpha$. That is,

$$\Pr(\tau_c \in C(\mathbf{Y})) \geq 1 - \alpha. \quad (16)$$

Proof. The set of all null hypotheses for which the null cannot be rejected is

$$C(\mathbf{y}) = \left\{ \tau^0 : T_{\min}(\mathbf{y}, \tau^0) - z_{\alpha/2} \leq 0 \right\} \cap \left\{ \tau^0 : T_{\max}(\mathbf{y}, \tau^0) + z_{\alpha/2} \geq 0 \right\}, \quad (17)$$

where we write $T_{\min}(\mathbf{y}, \tau^0) := \min_{g \in \mathcal{G}} \{T_g(\mathbf{y}, \tau^0)\}$ and $T_{\max}(\mathbf{y}, \tau^0) := \max_{g \in \mathcal{G}} \{T_g(\mathbf{y}, \tau^0)\}$ to make explicit the fact that the statistic depends on both data and the null hypothesis. For each possible τ^0 , we know from Proposition 1 that we have a level- α test. It follows that $C(\mathbf{y})$ is a confidence set with confidence coefficient $1 - \alpha$ (Theorem 9.2.2 in Casella and Berger 2002), hence equation (16) holds for the confidence set in equation (17).

To describe this confidence set, and to show that it is indeed an interval, it is helpful to note that by equation (8), $T_g(\mathbf{y}, \tau^0)$ is a decreasing linear function in τ^0 , meaning that $T_{\min}(\mathbf{y}, \tau^0)$ and $T_{\max}(\mathbf{y}, \tau^0)$ are strictly decreasing piecewise linear functions of τ^0 . Hence, we have

$$\left\{ \tau^0 : T_{\min}(\mathbf{y}, \tau^0) - z_{\alpha/2} \leq 0 \right\} = \left[\min_{g \in \mathcal{G}} \left\{ \tau^0 : T_g(\mathbf{y}, \tau^0) - z_{\alpha/2} = 0 \right\}, \infty \right), \quad (18)$$

$$\left\{ \tau^0 : T_{\max}(\mathbf{y}, \tau^0) + z_{\alpha/2} \geq 0 \right\} = \left(-\infty, \max_{g \in \mathcal{G}} \left\{ \tau^0 : T_g(\mathbf{y}, \tau^0) + z_{\alpha/2} = 0 \right\} \right], \quad (19)$$

The intersection of these two intervals yields the interval

$$C(\mathbf{y}) = \left[\min_{g \in \mathcal{G}} \left\{ \tau^0 : T_g(\mathbf{y}, \tau^0) - z_{\alpha/2} = 0 \right\}, \max_{g \in \mathcal{G}} \left\{ \tau^0 : T_g(\mathbf{y}, \tau^0) + z_{\alpha/2} = 0 \right\} \right] \quad (20)$$

By replacing $T_g(\mathbf{y}, \tau^0)$ with equation (8) we get the interval in equation (15). $\square$

A simple procedure for constructing the confidence interval is to form a confidence

interval for each double difference δ_g as $D_g \pm z_{\alpha/2}S_g$ and then take the end points of the union of all these confidence intervals.⁶

As shown in Section 3.1, the identification assumption implies bounds on the treatment effect. The bounds are consistently estimated by $\min_{g \in \mathcal{G}} D_g$ and $\max_{g \in \mathcal{G}} D_g$. Since, in addition, $\text{plim } S_g = 0$, we have that the confidence interval converges asymptotically to the bounds on the treatment effect given by equation (5).

3.4 Estimators

As discussed in Section 3.2, we need an estimator of δ_g , D_g , which has the property that it is asymptotically normal with a variance that can be consistently estimated. There are several such estimators that one can consider. With random sampling and i.i.d. observations, the most obvious candidate is the double difference in means

$$D_g = \bar{Y}_c - \bar{Y}_{c-1} - (\bar{Y}_g - \bar{Y}_{g-1}), \quad (21)$$

where $\bar{Y}_g = (1/N_g) \sum_{i=1}^n \mathbb{1}\{X_i = x_g\}Y_i$ is the sample mean of Y for units with $X = x_g$. Let $s_g^2 = (1/(N_g - 1)) \sum_{i=1}^n \mathbb{1}\{X_i = x_g\}(Y_i - \bar{Y}_g)^2$. A consistent estimator of $\mathbb{V}(D_g)$ is

$$S_g^2 = \begin{cases} \frac{1}{N_g} s_g^2 + \frac{1}{N_{g-1}} s_{g-1}^2 + \frac{1}{N_c} s_c^2 + \frac{1}{N_{c-1}} s_{c-1}^2 & \text{if } |g - c| > 1, \\ \frac{1}{N_c} s_c^2 + \frac{4}{N_{c-1}} s_{c-1}^2 + \frac{1}{N_{c-2}} s_{c-2}^2 & \text{if } g = c - 1, \\ \frac{1}{N_{c+1}} s_{c+1}^2 + \frac{4}{N_c} s_c^2 + \frac{1}{N_{c-1}} s_{c-1}^2 & \text{if } g = c + 1. \end{cases} \quad (22)$$

An alternative is to instead estimate the regression model

$$Y_i = \sum_{g=1}^G \gamma_g \mathbb{1}_{ig} + \varepsilon_i, \quad (23)$$

where $\mathbb{1}_{ig}$ is an indicator taking value of 1 if, for unit i , $X_i = x_g$, and 0 otherwise. The test can now be based on a series of Wald tests. Under this model, the estimator of δ_g is

$$D_g = \hat{\gamma}_c - \hat{\gamma}_{c-1} - (\hat{\gamma}_g - \hat{\gamma}_{g-1}). \quad (24)$$

which is analytically equivalent to the mean-difference estimator in equation (21). The

⁶The procedure for constructing the confidence interval has similarities to the sup-t method of constructing confidence bands for time series (Montiel Olea and Plagborg-Møller, 2019).

variance estimator would then be

$$S_g^2 = \mathbf{R}_g \hat{\Sigma} \mathbf{R}_g', \quad (25)$$

where $\hat{\Sigma}$ is the estimated covariance matrix and $\mathbf{R}_g$ is the $1 \times G$ matrix such that $\mathbf{R}_g \hat{\gamma} = \hat{\gamma}_c - \hat{\gamma}_{c-1} - (\hat{\gamma}_g - \hat{\gamma}_{g-1})$. If the usual heteroskedasticity-robust covariance matrix is used, this variance estimator is equivalent to the one in equation (22).⁷ However, other covariance matrices can also be used. For instance, suppose that individuals are observed several times. In that case, a cluster-robust covariance matrix, clustering at individual level, can be used.⁸

The regression implementation allows for the use of control variables, as well. Consider the model

$$Y_i = \sum_{g=1}^G \gamma_g \mathbb{1}_{ig} + \mathbf{Z}_i \beta + \eta_i, \quad (26)$$

where $\mathbf{Z}_i$ is a vector of control variables. Let $\tilde{\mu}(x_g) := \mathbb{E}(Y_i - \mathbf{Z}_i \beta)$. All results go through if we re-state Assumption 1 in terms of $\tilde{\mu}(x_g)$ and impose appropriate regularity conditions.

3.5 Bandwidth selection

In RD designs with continuous running variable, bandwidth selection plays an important role. Since identification relies on getting arbitrarily close to the cutoff, smaller bandwidths result in less bias, but reduction in sample size means increased variance. With a continuous running variable, bandwidth can be chosen optimally in a data-driven way (Cattaneo et al., 2020). In a discrete setting, optimal bandwidth selection continues to be feasible if one is willing to make assumptions about the behavior of the CEF in between support points (Kolesár and Rothe, 2018).

The situation is different in our setting. Assumption 1 is stated with explicit reference to a bandwidth, namely the closed interval $[x_1, x_G]$. Let us denote this assumption by $A1_{[1, G]}$. Now consider the interval $[a, b] \subseteq [x_a, x_b]$ and a modified Assumption $A1_{[a, b]}$ such that

⁷There can be slight differences depending on exactly how the robust covariance matrix is calculated. For instance, the covariance matrix HC0 uses a formula equivalent to equation (22), except that each term is multiplied with $(N_g - 1)/N_g$ (or, equivalently, $s_g^2 = 1/N_g \sum_{i=1}^{N_g} \mathbb{1}\{X_i = x_g\} (Y_i - \bar{Y}_g)^2$). The covariance matrix HC1 (which Stata uses as default in regression when the option `robust` is used) is equivalent to HC0, except that the whole expression is also multiplied with $n/(n - G)$.

⁸It is not possible to cluster at the running variable, since there is no variation in the data that will allow for estimating such intra-cluster correlations.

Assumption A1_[a,b]. $\min_{g \in \mathcal{G} \cap [a,b]} \{\mu(x_g) - \mu(x_{g-1})\} \leq \mu_0(x_c) - \mu(x_{c-1}) \leq \max_{g \in \mathcal{G} \cap [a,b]} \{\mu(x_g) - \mu(x_{g-1})\}$.

We have that A1_[a,b] implies A1_[1,G], while the opposite implication does not hold. In other words, narrowing the bandwidth in our case means strengthening the identification assumption.

A stronger identification assumption means a larger rejection region, and this increases statistical power. The tradeoff between power and strength of assumption is, in a sense, the reverse of the tradeoff in the continuous RD setting. In that case, as bandwidth decreases, the functional form imposed in the estimation becomes less restrictive while power decreases. Here, as bandwidth decreases, the assumption becomes more restrictive, while power increases. Note that this increase in power does not only hold in expectation: For any given sample, power is weakly increasing as bandwidth decreases.

Because the assumption involves the fundamentally unobserved outcome under no treatment for treated units, there is no direct data-driven way of choosing bandwidth. Rather, the choice of bandwidth has to be made based on subject-matter knowledge of the application at hand. This gives researchers many degrees of freedom and they may be tempted to set a very small bandwidth to increase the chance of finding statistically significant effects. As the spirit of the methodology developed in this paper is to be conservative, we recommend to always show results for the largest possible bandwidth. But since power may be poor in such cases—especially with many support points—these results can be complemented by showing results for a range of bandwidths. It is not unusual for extreme values of the running variable to contain much fewer observations, which means power will be much lower when these values of the running variable are included. In such cases, it is reasonable for the researcher to argue that a smaller bandwidth is preferable, but by showing results for a range of bandwidths, the problem of selective reporting is mitigated.

3.6 Illustration using synthetic data

To illustrate our test procedure, we use an example based on synthetic data, displayed in Table 1, with $G = 14$ and $c = 8$. Columns 2 and 3 show the mean and standard error of the outcome at each point of the running variable, with Figure 2a illustrating the mean and confidence interval. In columns 4 and 5, the double difference D_g and its standard error are shown. Finally, the last column shows the t -statistics.

Based on these data, $T_{\min}(\mathbf{y}) = 0.53$ and $T_{\max}(\mathbf{y}) = 3.53$, so that $\mathcal{T}(\mathbf{y}) = 0.53$, and $p(\mathbf{y}) = 2(1 - \Phi(0.53)) = 0.60$. We can form a 95-percent confidence interval for τ_c by first forming confidence intervals for each $\delta_g, g \in \mathcal{G}$, that is, $D_g \pm 1.96 \cdot S_g$. The smallest

x_g (1)	$\bar{Y}_g$ (2)	$s_g/\sqrt{N_g}$ (3)	D_g (4)	S_g (5)	T_g (6)
1	0.28	0.077			
2	0.13	0.065	0.41	0.120	3.41
3	0.32	0.053	0.08	0.107	0.70
4	0.29	0.051	0.30	0.099	3.00
5	0.41	0.053	0.15	0.099	1.48
6	0.43	0.049	0.25	0.097	2.56
7	0.49	0.045	0.21	0.113	1.82
8	0.75	0.048	0	0	
9	0.82	0.049	0.20	0.117	1.67
10	0.75	0.048	0.34	0.095	3.53
11	0.79	0.054	0.22	0.098	2.26
12	0.92	0.056	0.14	0.102	1.41
13	0.91	0.065	0.27	0.108	2.49
14	1.11	0.079	0.06	0.121	0.53

Table 1: Synthetic data example

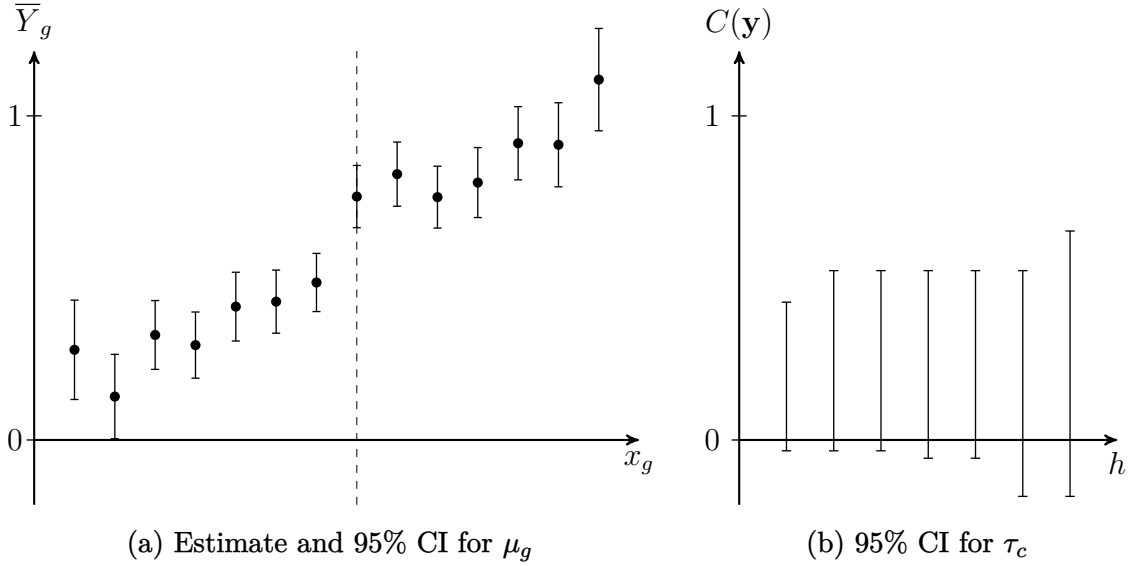


Figure 2: Illustration of synthetic data

value of all these intervals is for $g = 14$ with a value of $0.06 - 1.96 \cdot 0.121 = -0.17$. Similarly, the largest point is for $g = 2$ with a value of $0.41 + 1.96 \cdot 0.120 = 0.64$. Therefore, the 95-percent confidence interval goes from -0.17 to 0.64 . Finally, the bounds on the treatment effect are estimated to be $[\min_{g \in \mathcal{G}}\{D_g\}, \max_{g \in \mathcal{G}}\{D_g\}] = [0.06, 0.41]$.

Just as with regular RD, the result will depend on the chosen bandwidth. Hence, we recommend displaying confidence intervals for a range of bandwidths. Figure 2b does so by displaying confidence intervals for data in the interval $[c - h, c + h]$, with h ranging from 1 to 7. As discussed in Section 3.5, the confidence interval is guaranteed to be weakly increasing as the bandwidth increases. We also see that the end points of the confidence interval may remain unchanged with an increase in bandwidth.

3.7 Running variable with unequal spacing

Within our framework, the running variable can be measured on any scale which can be ordered. However, as discussed above, when the running variable is measured on an interval or ratio scale (that is, the quantitative difference between two points is meaningful) and the running variable is not equally spaced, then it will in general be preferable to state the identification assumption in terms of normalized differences. The assumption would then be

Assumption 3. $\min_{g \in \mathcal{G}} \left\{ \frac{\mu(x_g) - \mu(x_{g-1})}{x_g - x_{g-1}} \right\} \leq \frac{\mu_0(x_c) - \mu(x_{c-1})}{x_c - x_{c-1}} \leq \max_{g \in \mathcal{G}} \left\{ \frac{\mu(x_g) - \mu(x_{g-1})}{x_g - x_{g-1}} \right\}.$

The double-difference estimator is then

$$D_g = \bar{Y}_c - \bar{Y}_{c-1} - \varphi_g (\bar{Y}_g - \bar{Y}_{g-1}), \quad \varphi_g := \frac{x_c - x_{c-1}}{x_g - x_{g-1}} \quad (27)$$

with variance estimator

$$S_g^2 = \begin{cases} \frac{s_c^2}{N_c} + \frac{s_{c-1}^2}{N_{c-1}} + \varphi_g^2 \left(\frac{s_g^2}{N_g} + \frac{s_{g-1}^2}{N_{g-1}} \right) & \text{if } |g - c| > 1, \\ \frac{s_c^2}{N_c} + (1 + \varphi_{c-1})^2 \frac{s_{c-1}^2}{N_{c-1}} + \varphi_{c-1}^2 \frac{s_{c-2}^2}{N_{c-2}}, & \text{if } g = c - 1, \\ (1 + \varphi_{c+1})^2 \frac{s_c^2}{N_c} + \frac{s_{c-1}^2}{N_{c-1}} + \varphi_{c+1}^2 \frac{s_{c+1}^2}{N_{c+1}}, & \text{if } g = c + 1. \end{cases} \quad (28)$$

3.8 Multiple cutoffs

In some settings, several treatments are assigned based on distinct cutoffs on a common running variable. For instance, reported test scores may be a step function of raw points, or there may be successive reforms such as increases of the minimum school-leaving age. The applied researcher faces two options. First, they may perform the analysis separately for each cutoff, using standard tools to adjust for multiple comparisons. Second, they may stack the data to estimate a single target parameter, yielding increased power and avoiding the multiple comparison issue. Here we focus on the second approach.

As before, there are G support points, but there are now $K < G$ treatments indexed by $k = 1, 2, \dots, K$, with corresponding cutoff points $x_{c(k)}$ such that treatment is assigned according to the function

$$W_i = \sum_{k=1}^K k \mathbb{1}\{x_{c(k)} \leq X_i < x_{c(k+1)}\}, \quad (29)$$

where $x_{c(K+1)} > x_G$. The potential outcome under treatment k is $Y_i(k)$ and we define $\mu_k(x_g) := \mathbb{E}(Y_i(k)|X_i = x_g)$. Actual outcomes are $Y_i = Y_i(W_i)$ and group means are $\mu(x_g) := \mathbb{E}(Y_i|X_i = x_g)$. The treatment effect at cutoff k is $\tau_{c(k)} := \mu(x_{c(k)}) - \mu_{k-1}(x_{c(k)})$. Define the weights $\omega_k := p_{c(k)} / \left(\sum_{k'=1}^K p_{c(k')}\right)$. The target parameter when stacking is

$$\tau^* := \sum_{k=1}^K \omega_k \tau_{c(k)}, \quad (30)$$

a weighted average of the average treatment effects at the cutoffs.

We focus on the case of an equally-spaced running variable, with the additional refinement of unit spacing, $x_g - x_{g-1} = 1$, $\forall g > 1$. The refinement is innocuous but considerably eases notation. We discuss unequal spacing below.

Let us now stack the data in the following way. Around each cutoff, select observations with $x_g \in I_k := [x_{c(k)} - h^-, x_{c(k)} + h^+]$, where h^-, h^+ are positive integers and it must be ensured that I_k does not contain another change in treatment, $x_{c(k)} - h^- \geq x_{c(k-1)}$ and $x_{c(k)} + h^+ < x_{c(k+1)}$. For simplicity, we restrict the bandwidth to be common for all cutoffs.

Define the set $\mathcal{L} = \{-h^-, -h^- + 1, \dots, -1, 1, \dots, h^+ - 1, h^+\}$. Define the weights $\omega_{k,l} := p_{c(k)+l} / \left(\sum_{k'=1}^K p_{c(k')+l}\right)$. For l steps away from each cutoff, the stacked outcome means are defined as

$$\mu^*(l) := \sum_{k=1}^K \omega_{k,l} \mu(x_{c(k)+l}), \quad l \in \mathcal{L} \cup \{0\}. \quad (31)$$

Furthermore, define the average untreated outcome for units at any cutoff,

$$\mu_0^*(0) := \sum_{k=1}^K \omega_k \mu_{k-1}(x_{c(k)}). \quad (32)$$

The identification assumption with stacked data becomes

Assumption 2. $\min_{l \in \mathcal{L}} \{\mu^*(l) - \mu^*(l-1)\} \leq \mu_0^*(0) - \mu^*(-1) \leq \max_{l \in \mathcal{L}} \{\mu^*(l) - \mu^*(l-1)\}.$

We can now test the null that $\tau^* = \tau^0$, as well as forming a confidence interval for τ^* using the same approach as before, but with stacked data. The stacked sample mean at point l is

$$\bar{Y}_l^* = \sum_{k=1}^K \Omega_{k,l} \bar{Y}_{c(k)+l}, \quad l \in \mathcal{L} \cup \{0\}, \quad (33)$$

where $\Omega_{k,l} := N_{c(k)+l} / \left(\sum_{k'=1}^K N_{c(k')+l} \right)$, and the double-difference estimator according to equation (21) is

$$D_l = \bar{Y}_0^* - \bar{Y}_{-1}^* - (\bar{Y}_l^* - \bar{Y}_{l-1}^*), \quad l \in \mathcal{L}. \quad (34)$$

If there is no overlap of the bandwidth window between cutoffs, equation (22) with stacked data can be used. However, if bandwidth windows overlap, such that an observation can be on both sides of cutoffs, then that needs to be taken into account. The easiest way to do this is to use OLS to estimate the model

$$Y_{il} = \sum_{l=-h^-}^{h^+} \gamma_l \mathbb{1}_{il} + \varepsilon_{il}, \quad (35)$$

where $\mathbb{1}_{il}$ is an indicator if unit i is in the bandwidth window of a given cutoff. Hence, if a unit is part of two different bandwidth windows, there will be two observations, and the cluster-robust covariance matrix, clustered at the unit level, should be used. Everything else proceeds in the exact same way as outlined in Section 3.4.

The relationship between Assumption 2 on the one hand, and Assumption 1 holding at each cutoff on the other, requires comment. We continue to focus on the case of a common bandwidth. It turns out that neither does Assumption 1 imply Assumption 2, nor does Assumption 2 imply Assumption 1. Assumption 2 may hold even if Assumption 1 is violated at all cutoffs, for instance because violations average out. On the other hand, Assumption 2 may not hold even if Assumption 1 is satisfied at each cutoff, due to different distributions of data across support points at different cutoffs—a version of Simpson’s paradox.

An alternative is to replace $\omega_{k,l}$ with ω_k in equation (31), meaning that each support point gets weighted equally around a given cutoff.⁹ Under this modification, if Assumption 1 holds at every cutoff, then Assumption 2 holds. The converse still does not hold. Hence, in that case, Assumption 2 is weaker than Assumption 1 holding at each cutoff.

An equally spaced running variable is common, with examples including test scores or rankings. With a single cutoff, we showed in Section 3.7 how non-equal spacing is easily handled by our approach. With multiple cutoffs, it is not obvious how to stack the data while adjusting for the unequal spacing, or how to formulate the identification assumption. To get around this issue, one may replace the running variable by its rank, a solution that is always feasible. The identification assumption must then be stated with respect to the rank.

4 Monte Carlo experiments

In this section we report results from a simulation study to compare the performance of our double-differences (DD) approach to that of the two approaches proposed by Kolesár and Rothe (2018), one based on a bound on the second derivative (BSD) and the other based on the assumption that the estimated CEF fits no worse at the cutoff than anywhere else (bounded misspecification error, BME). We introduce our setup in Section 4.1. We discuss theoretical properties of the various estimators, given the setup, in Section 4.2. Section 4.3 contains the results.

4.1 Setup

There are five dimensions along which we vary the specification of the simulations. First, the number of support points. Second, the degree of nonlinearity. Third, the steepness of the CEF around the cutoff. Fourth, sample size. And last, the size of the treatment effect.

The data generating process is

$$Y_i = X_i + \lambda \sin(\pi(X_i + \zeta_j)) + \tau W_i + \varepsilon_i, \quad \zeta_j := \frac{1 + 2j}{G - 1}, \quad (36)$$

with i.i.d. errors $\varepsilon_i \sim \mathcal{N}(0, 0.1)$ and W_i taking the value one whenever $X_i \geq x_c$ and zero otherwise. There are $G \in \{11, 21\}$ support points equally spaced on the interval $[-1, 1]$. The probability of unit i being on support point x_g is a constant,

⁹To implement this approach, weighted least squares must be used when estimating equation (35) with weights $N_{c(k)}/N_{c(k)+l}$.

$\Pr(X_i = x_g) = 1/G$. The cutoff is located at $x_c = 0$.

The CEF is a sine wave around an upward sloping line. The parameter λ controls how much the function deviates from linearity, where we consider values $\lambda \in \{0, 0.05, 0.15, 0.25\}$. The phase shifter ζ_j , for $j \in \{1, 2, \dots, G-1\}$, creates variation in the relative steepness of the CEF around the cutoff. The first differences of the CEF follow a cyclical pattern, and the support of the running variable always contains a full cycle. See Figures A1 and A2 for illustrations. Our CEF is similar to that used in the Monte Carlo experiments of Kolesár and Rothe (2018).

We set the sample size at $n \in \{500; 1,000; 5,000; 25,000\}$. The normalized treatment effect takes on eleven distinct values $\tau/\sqrt{\mathbb{V}(\varepsilon)} \in \{0, .1, .2, \dots, .9, 1\}$.

There are 5,280 different specifications in total (for combinations of λ, j, n, τ , there are $4 \cdot 10 \cdot 4 \cdot 11 = 1,760$ specifications when $G = 11$ and $4 \cdot 20 \cdot 4 \cdot 11 = 3,520$ specification when $G = 21$). For each of these 5,280 specifications we perform 10,000 replications. X and ε and are stochastic; hence, their realized values differ for each replication.

We perform all simulations using Julia (julialang.org) to speed up the simulations. We have cross-checked our implementation of the BME and BSD estimators with the implementations in the Stata package `rdhonest` and the R package `RDHonest`.

4.2 Theoretical properties

Before presenting our simulation results we discuss two theoretical issues. First, we examine whether the identification assumption for each approach is satisfied. And second, we characterize asymptotic confidence intervals for each approach.

Assumption 1 underlying the DD approach is satisfied in all cases except when the minimum or maximum of the first-differenced sine function is located at the cutoff: namely, for $\lambda > 0$, when $G = 11$ and $j \in \{5, 10\}$, and when $G = 21$ and $j \in \{10, 20\}$. The reason we add $\pi/(G-1)$ inside the sine function is precisely to create such violations. Without the addition there would be ties, and Assumption 1 would always be satisfied.

In the case of BME, the identification assumption is that the estimated CEF fits no worse at the cutoff than elsewhere within the chosen bandwidth. Following Kolesár and Rothe (2018), we fit regression lines on each side of the cutoff using a uniform kernel. This means that the identification assumption will often be violated when $\lambda > 0$, because a linear approximation to a nonlinear function such as the one used here tends to fit worst at the edges, exactly where the cutoff is located.¹⁰

¹⁰In our setting, the identification assumption always holds to the right of the cutoff, but to the left of the cutoff it is always violated when $G = 11$, regardless of phase shift. For $G = 21$, it is violated to the left of the cutoff in all cases except when $j \in \{2, 3, 4, 12, 13, 14\}$.

For the BSD estimator we also use a uniform kernel and choose the true bound $K = \lambda\pi^2$, so that the identification assumption underlying BSD is always satisfied. Recall that in practice, the applied researcher would not know this bound precisely and might either make a more conservative choice or a choice which violates the assumption for the BSD estimator. In the appendix, we instead show results where K is chosen in the data-driven manner suggested by Armstrong and Kolesár (2020).

Because none of the three approaches assume a specific functional form of the CEF, the confidence intervals do not in general converge to a point asymptotically, but instead converge to finite intervals. The DD estimator converges to the bounds on the treatment effect implied by the data generating process, $[\min_{g \in \mathcal{G}} \{\delta_g\}, \max_{g \in \mathcal{G}} \{\delta_g\}]$. Given equation (36), the length of this asymptotic confidence interval is¹¹

$$L_{DD} = \max_{g \in \mathcal{G}} \{\delta_g\} - \min_{g \in \mathcal{G}} \{\delta_g\} = 4\lambda \sin\left(\frac{\pi}{G-1}\right). \quad (37)$$

For the BSD estimator, the bound has the following length (assuming the CEF is defined over all reals in the interval $[-1, 1]$):

$$L_{BSD} = \lambda\pi^2 \frac{G^2 + 3}{3(G-1)^2}. \quad (38)$$

Comparing equations (37) and (38), for any $G \geq 3$ and $\lambda \neq 0$, we have $L_{DD}/L_{BSD} < 1$. In fact, the ratio is at most around 1/2 at $G = 4$. Hence, under the present data generating process, the DD estimator always has a tighter asymptotic bound than the BSD estimator, except when $\lambda = 0$ in which case both converge to bounds of length zero. We can also see that both bounds have length that gets narrower as G increases. While we focus on settings with finite, fixed, G in this paper, we note that the length of the bound converges to zero for the DD estimator as G tends to infinity.¹² However, that is not the case for the BSD estimator.

It is more challenging to derive asymptotic bounds for the BME estimator in general, as they will depend on phase shift. However, it is straightforward to compute these bounds for each value of j . We will find that the asymptotic bounds of the BME estimator tend to be of greater length than those of the DD estimator.

The fact that our DD estimator has the smallest asymptotic bounds means that there exists a sample size above which the DD estimator has the smallest confidence interval in expectation, under the nonlinear specifications considered. However, in smaller samples this is not guaranteed to be the case, as we will see shortly.

¹¹This length is for the cases where the assumption is fulfilled. If it is not, the length will be slightly smaller.

¹²In fact, this result holds under any CEF which converges to a continuous function as $G \rightarrow \infty$, as long as X is defined over a fixed range.

4.3 Monte Carlo results

We begin with the case $\tau = 0$ and study precision and coverage properties for the three estimators. For precision, we follow Kolesár and Rothe (2018) and report normalized standard errors defined as $(C^+ - C^-)/(2 \cdot 1.96)$, where C^- and C^+ are the lower and upper points of the confidence intervals. Panel A of Table 2 reports these normalized standard errors averaged across values of the phase shifter, for all choices of support points, sample size, and degree of nonlinearity.

Panel A, columns 1–3 of Table 2 reveal that under a linear model, BSD is the most precise, while DD and BME have similar precision but much lower than that of BSD. However, here we have set the bound on the second derivative for BSD to zero. While this is correct given the model, in practice an applied researcher will always choose a positive bound. Asymptotically, DD is efficient in the sense that the standard error goes to zero. In finite samples, however, it is conservative, because our test procedure makes multiple comparisons (even though Assumption 1 holds with equality).

When the model is nonlinear (columns 4–12 in panel A of Table 2), the relative performance of DD and BME improves and they outperform BSD eventually as sample size or the degree of nonlinearity increases. We also see that DD outperforms BME at sufficient sample sizes and degree of nonlinearity, especially when $G = 21$. Given the theoretical discussion above, this is expected as far as DD and BSD are concerned. The lines where the sample size is indicated to be infinite contain the normalized standard errors based on the asymptotic bounds. As we see, DD also tends to have tighter asymptotic bounds than BME.

Figure 3 summarizes to what extent, and under what conditions, our DD estimator outperforms the available alternatives—BSD and BME—in terms of precision. The figure is constructed as follows. Let $\nu_s^{DD}, \nu_s^{BSD}, \nu_s^{BME}$ denote the normalized standard errors for the respective estimators and specification s . First, for each s we calculate the relative standard error as $\nu_s^{DD} - \min\{\nu_s^{BSD}, \nu_s^{BME}\}$. When this number is negative, DD outperforms the alternatives in terms of precision. In cases where the assumption of an estimator fails (10 percent of cases for DD, 60 percent of cases for BME and in no case for BSD), for the purpose of this figure, we let the standard error be infinite. Second, we plot the empirical CDF of the relative standard error across the 1,320 specifications for each value of λ . The fraction of specifications for which DD dominates the alternatives equals the intercept of each line.

Figure 3 reveals that DD offers higher precision in the vast majority of cases for moderate to highly nonlinear CEFs: With $\lambda = 0.15$ it outperforms both BME and BSD in 68 percent of cases and when $\lambda = 0.25$ it does so in 83 percent of cases. As expected, in the linear case, $\lambda = 0$, DD is outperformed by the alternatives, namely BSD. This is because we set $K = 0$, which essentially reduced BSD to local linear

G		$\lambda = 0.0$				$\lambda = 0.05$				$\lambda = 0.15$				$\lambda = 0.25$					
		DD		BME		BSD		DD		BME		BSD		DD		BME		BSD	
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)						
Panel A: Normalized standard error, averaged over values of the phase shifter, ζ_j																			
11	500	.154	.143	.060	.158	.146	.102	.177	.164	.207	.202	.190	.311						
11	1,000	.108	.100	.042	.113	.104	.088	.135	.125	.192	.162	.154	.296						
11	5,000	.048	.044	.019	.056	.051	.068	.083	.079	.172	.112	.111	.276						
11	25,000	.022	.020	.008	.032	.030	.059	.062	.061	.163	.092	.094	.267						
11	∞	.000	.000	.000	.015	.016	.052	.046	.049	.156	.077	.082	.260						
21	500	.230	.223	.058	.230	.225	.095	.235	.240	.188	.244	.265	.282						
21	1,000	.160	.154	.041	.161	.157	.081	.167	.176	.174	.178	.205	.267						
21	5,000	.070	.068	.018	.072	.073	.062	.084	.103	.155	.098	.142	.248						
21	25,000	.031	.030	.008	.035	.040	.053	.049	.079	.147	.064	.122	.240						
21	∞	.000	.000	.000	.008	.022	.047	.024	.065	.140	.040	.108	.233						
Panel B: Coverage of 95% CI, averaged over values of the phase shifter, ζ_j																			
11	500	1.000	1.000	.949	1.000	.999	.990	.999	.973	.999	.997	.850	1.000						
11	1,000	1.000	1.000	.949	1.000	.998	.995	.998	.902	1.000	.996	.640	1.000						
11	5,000	1.000	1.000	.950	.999	.963	1.000	.994	.453	1.000	.988	.348	1.000						
11	25,000	1.000	1.000	.950	.996	.615	1.000	.980	.300	1.000	.938	.213	1.000						
11	∞	1.000	1.000	.950	.800	.200	1.000	.800	.200	1.000	.800	.200	1.000						
21	500	1.000	1.000	.948	1.000	1.000	.989	1.000	1.000	.999	1.000	.999	1.000						
21	1,000	1.000	1.000	.950	1.000	1.000	.994	1.000	.999	1.000	1.000	.989	1.000						
21	5,000	1.000	1.000	.950	1.000	1.000	.999	1.000	.951	1.000	1.000	.776	1.000						
21	25,000	1.000	1.000	.950	1.000	.988	1.000	1.000	.647	1.000	.999	.519	1.000						
21	∞	1.000	1.000	.950	.900	.400	1.000	.900	.400	1.000	.900	.400	1.000						

Table 2: Normalized standard errors and coverage of 95-percent confidence intervals

regression, resulting in efficient inference. In practice, when using BSD, K would never be set to zero.

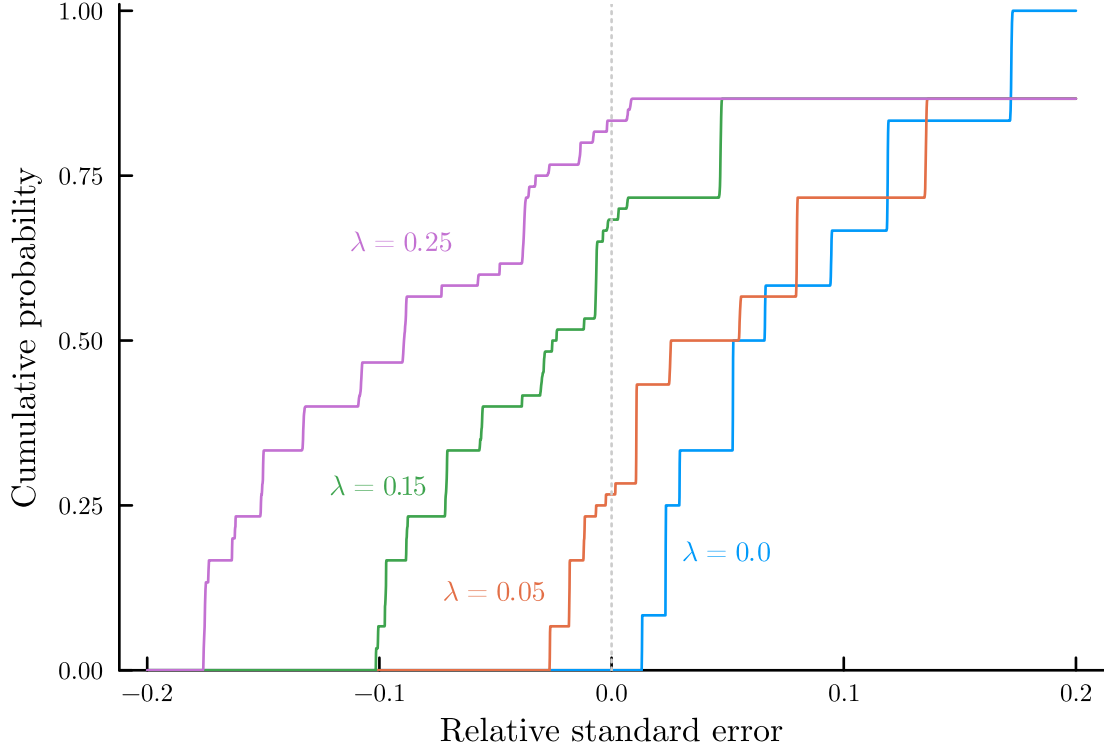


Figure 3: Empirical CDF of relative standard error from Monte Carlo simulations

Notes: The figure shows the empirical CDF of the the standard error of DD minus the minimum of the standard error of BSD and BME for each value of λ . The distributions are based on 1,320 different specifications for each value of λ .

Next, we turn to coverage properties reported in Panel B of Table 2. In almost all scenarios, both DD and BSD have conservative coverage on average. Recall, however, that Assumption 1 is violated in a small number of cases. This is reflected in the rows where the sample size is indicated to be infinite, which display the theoretical bounds averaged across phase shifts. In the case of DD, this average equals the fraction of cases where the identification assumption is satisfied. This is because when the assumption holds, the asymptotic bounds necessarily contain the true effect $\tau = 0$, whereas they inevitably exclude the true effect when the assumption is violated. In that sense, the bound is asymptotically the most efficient bound given the assumption. This is in contrast to the BME estimator where there are specific phase shifts when asymptotic coverage is unity despite the assumption failing: namely, when $G = 11$ and $j \in \{2, 7\}$, and when $G = 21$ and $j \in \{5, 15\}$.

Panel B of Table 2 also suggests that violation of the respective identification assumptions are more serious in the case of BME, which often under-covers. In fact,

when we inspect coverage as a function of phase shifts (Figure A3 in the appendix), we find that coverage of BME drops below 95 percent in many instances and even hits zero in some cases for the largest sample size. For the DD estimator, coverage drops below 95 percent only in a few instances for $G = 11$ when the CEF is nonlinear, but stays conservative in all other cases. To be clear, the discussion here concerns situations when the identification assumptions do not hold. When they do hold, both DD and BME have conservative coverage.

In Tables A1 and A2 in the appendix, we add results for the BSD estimator when K is chosen in the data-driven manner suggested by Armstrong and Kolesár (2020). We also show results from the standard approach of using local linear regression with robust standard errors, both with uniform and triangular kernel. Finally, we also reproduce the results from Table 2 for ease of comparison. With data-driven K , the BSD standard errors increase dramatically and are several times—and up to an order of magnitude—larger compared to the other estimators. In contrast, the local linear regression estimators have very small standard errors as expected, but the coverage is very poor as soon as nonlinearity is introduced, even with relatively modest degree of nonlinearity and with the use of a triangular kernel. Hence, in the setting we are interested in with relatively few support points, but many observations at each support point, we conclude that neither the local linear approach, nor the BSD approach with data-driven K , is suitable.

We study power as the rate of rejection when the significance level is five percent and the size of the treatment effect varies as described above. The results are displayed in Figure A4 in the appendix, where we plot rejection rates for the three different estimators as a function of the normalized treatment effect, $\tau / \mathbb{V}(\varepsilon)$, separately for the different sample sizes, degree of nonlinearity, and number of support points.

We first focus on DD and BSD. Here the findings largely mirror the discussion of precision above. DD tends to be relatively more powerful compared to BSD in nonlinear models as the sample size and degree of nonlinearity increase, as well as for larger effect sizes.

DD and BME have similar power when the CEF is linear ($\lambda = 0$). When the CEF is nonlinear, the comparison is complicated by the more frequent violations of the identification assumption of BME, and its greater sensitivity to such violations. BME often appears to have greater power, but this is because it has incorrect size. The size distortions are clearly visible where the BME rejection rates have an intercept well above five percent, and in some cases even decrease locally in effect size—meaning that the probability of rejecting the null in case of a zero treatment effect is greater than the probability of rejecting it when the treatment effect is non-zero.

To sum up, our DD approach compares favorably to alternative estimators when

the CEF is nonlinear. Although BSD is sometimes more powerful in our simulations, this happens because we set K to its true value, which is unknown in practice. As for BME, it has similar power to DD when identification assumptions hold, but violations occur more often in the case of BME, and size distortions are more severe than when the identification assumption of DD is violated. We suspect that this finding may generalize beyond our Monte Carlo experiments, due to the tendency of the linear fit to perform worst at the cutoff. BME may perform better with a triangular kernel, but implementation is not obvious.¹³

5 Applications

5.1 SAT scores and college enrollment

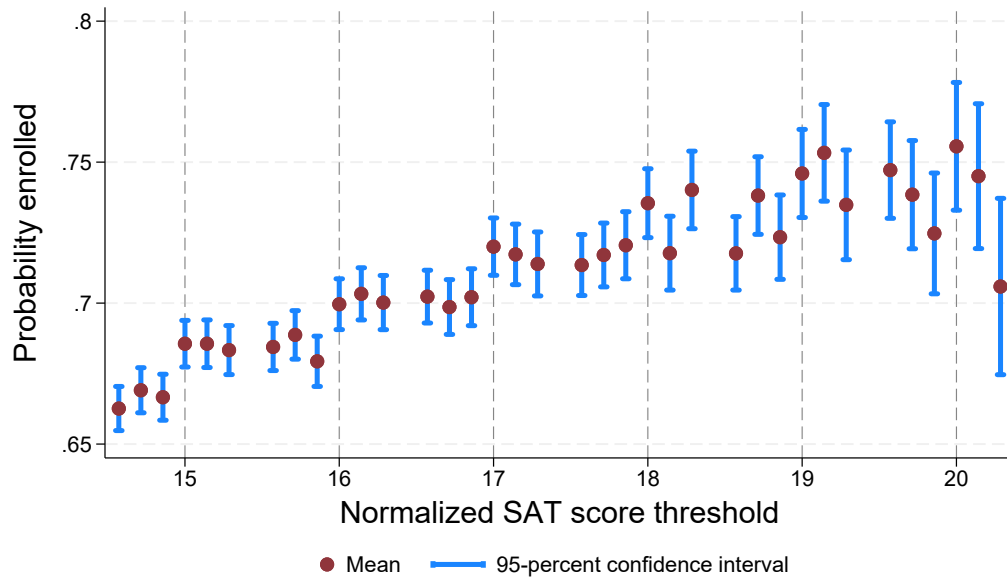
Graetz et al. (2025) use the mapping from raw to reported scores on the Swedish SAT to estimate the effect of higher SAT scores on college enrollment and other outcomes. Reported scores are on a 20-point scale. Test takers answer about 100 multiple choice questions. The number of correct answers is a raw score that is mapped to the final reported score based on cutoffs determined after the test. The cutoffs are set in such a way that the distribution of final scores approximates a normal distribution.

There are typically only a few steps on the raw score between consecutive final-score thresholds. A regression discontinuity design applied to this context thus features a discrete running variable with few support points. Graetz et al. (2025) use parametric methods for estimation and inference, and use a permutation-based approach in a robustness check. Here we apply our nonparametric method to their setting.

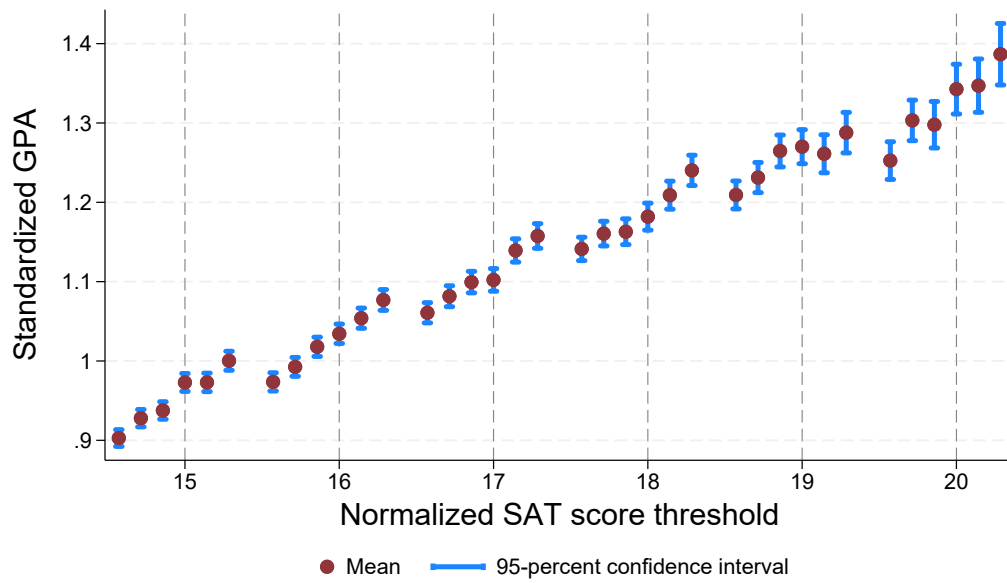
Figure 4a plots the fraction enrolled in college one year after the test, along with its 95-percent confidence interval, over three support points on each side of the six upper thresholds. There appear to be some unusual jumps at the cutoffs, at the same time it does not seem innocuous to impose linear regression functions. Figure 4b shows the corresponding plot for compulsory school GPA, which serves as a placebo outcome (GPA is set before one is allowed to take the SAT).

We apply our approach separately to each of the six score thresholds (15–20), as well as to the stacked data. Results are displayed in Table 3. For each of the six thresholds, as well as for the stacked data, columns 1 and 2 display the minimum and maximum t -statistics, respectively. Column 3 shows the test statistic, the closest distance between zero and the interval spanned by the t -statistics, and column 4 displays the associated p -values. The remaining two columns contain upper and lower

¹³The only implementation of the BME estimator that we are aware of, the R package `RDHonest`, only allows for a uniform kernel.



(a) College enrollment a year after taking the SAT



(b) Compulsory school GPA

Figure 4: College enrollment and compulsory school GPA against SAT scores

Notes: The figure shows mean college enrollment two years after taking the Swedish SAT and compulsory school (year-9) GPA, along with 95-percent confidence intervals, by raw score grouped around normalized score thresholds. Data are from Graetz et al. (2025).

bounds of 95-percent confidence intervals.

	$\min\{T_g\}$ (1)	$\max\{T_g\}$ (2)	$\mathcal{T}$ (3)	p -value (4)	C^- (5)	C^+ (6)
<i>Panel A. Enrolled</i>						
By threshold						
15	1.52	2.48	1.52	0.129	-0.004	0.041
16	1.47	2.67	1.47	0.142	-0.006	0.051
17	1.15	2.15	1.15	0.249	-0.010	0.046
18	-0.58	2.12	0.00	1.000	-0.033	0.063
19	0.14	2.38	0.14	0.886	-0.027	0.075
20	1.44	2.69	1.44	0.149	-0.015	0.121
Stacked	3.50	4.50	3.50	0.000	0.007	0.034
Stacked, reweighted	3.43	4.49	3.43	0.001	0.006	0.034
<i>Panel B. GPA</i>						
By threshold						
15	0.69	2.50	0.69	0.488	-0.015	0.063
16	-0.59	-0.20	0.20	0.842	-0.039	0.027
17	-1.93	-0.88	0.88	0.377	-0.069	0.018
18	-0.68	0.82	0.00	1.000	-0.050	0.057
19	-1.11	0.53	0.00	1.000	-0.077	0.067
20	-0.21	1.39	0.00	1.000	-0.062	0.122
Stacked	-0.38	0.91	0.00	1.000	-0.015	0.022
Stacked, reweighted	-0.79	0.66	0.00	1.000	-0.018	0.020

Table 3: Tests and confidence intervals for the effects of SAT score on college enrollment and compulsory schooling GPA

Panel A of Table 3 shows that the change at the cutoff is the largest compared to changes away from the cutoff for all thresholds except for 18 (as $\min\{T_g\}$ is negative). However, the null cannot be rejected at any given cutoff and, consequently, the confidence intervals all cover zero. We also report results from two stacked specifications. In the first, each support point gets weight proportional to its sample size, whereas in the reweighted version, the weight is enforced to be the same for all support points within the bandwidth at a given threshold (see Section 3.8). Here we see that when data are pooled, there is a clear positive effect of receiving a higher normalized SAT score on college enrollment with a 95-percent confidence interval ranging from 0.6–0.7 to 3.4 percentage points. This result is consistent with the results from the parametric

approach used by Graetz et al. (2025).¹⁴

Panel B of Table 3 instead shows the results for the placebo outcome of compulsory GPA. Here we see that there is no evidence of a treatment effect at any threshold with large p -values. Reassuringly, the same is true also for the stacked versions.

As an example, Figure 5a illustrates our procedure for the threshold around the normalized score of 16. There does appear to be an unusual jump at the cutoff in the left panel. This is confirmed by plotting the double differences in the right panel. The estimated bounds of the treatment effect are 1.6 and 3 percentage points, and the 95-percent confidence interval only just includes zero. Figure 5b shows a qualitatively similar pattern for the stacked data. However, the confidence interval is now substantially narrower, as are the estimated bounds of the treatment effect.

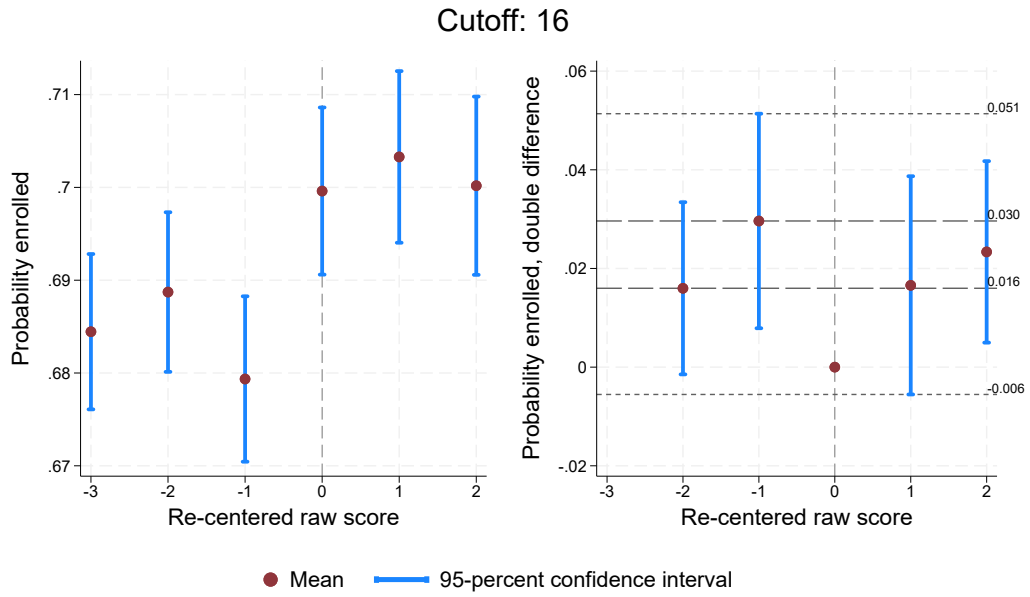
Not surprisingly, SAT raw scores are strongly positively correlated with predetermined variables, such as the compulsory GPA (Figure 4b). This is why a local randomization approach, for instance only using a single point on each side of the cutoff, is invalid. In contrast, our approach is based on double differences and will not be contaminated by a systematic relationship between raw scores and unobserved traits, as long as Assumption 1 holds.

5.2 Minimum school-leaving age and earnings

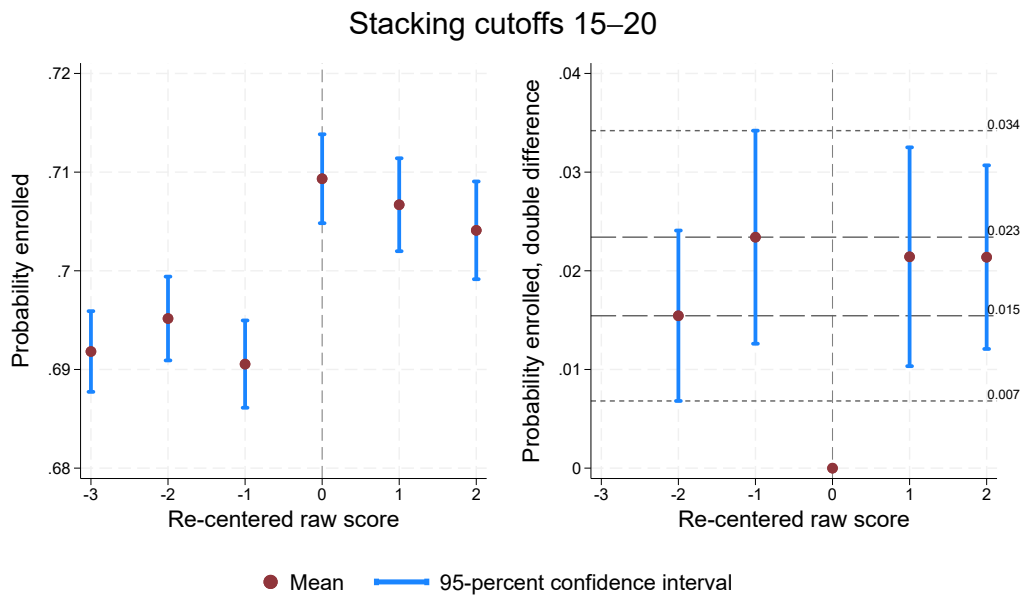
In April 1947, the UK government changed the minimum age at which youths were allowed to leave school from 14 to 15. Oreopoulos (2006, 2008) estimates the effect of the reform on earnings using data from the UK General Household Surveys by employing a sharp RD design. Oreopoulos (2006, 2008) uses parametric methods and clusters standard errors at the level of the running variable. Here we implement our approach in that setting and compare our results to those from Kolesár and Rothe (2018), where we contrast with the BME and BSD approaches, as well as the standard local polynomial regression approach with Eicker-Huber-White (EHW) robust standard errors.

Table 4 shows the results, where all results except our DD estimator are replications of figures from Table 3 in Kolesár and Rothe (2018). For EHW and BME, Kolesár and Rothe (2018) show results for three different bandwidths (3, 6 and ∞), which we replicate here. For BSD, Kolesár and Rothe (2018) determine bandwidth in a data-driven way. For the intermediate choices of the bound $K = 0.02$ and $K =$

¹⁴Stacking of data in this application involves using individuals repeatedly, which is why Graetz et al. (2025) clustered at the individual level when calculating standard errors for their parametric specifications. As we no longer have access to the micro data, we are unable to estimate the cluster-robust covariance matrix. Hence, the confidence intervals for the pooled treatment effects we present here are likely too narrow.



(a) A single cutoff



(b) Stacking cutoffs

Figure 5: DD test for the effects of higher SAT scores on college enrollment

Notes: In the right panels, short-dashed lines mark 95-percent confidence intervals for the treatment effect. Long-dashed lines mark the estimated bounds on the treatment effect.

0.04 that we replicate here, the bandwidths are 3 and 2. In Table 4, we report DD confidence intervals and p -values corresponding to bandwidths $(2, 3, 6, \infty)$, and for all other bandwidths in between in Figure 6a.

	(1)	(2)	(3)	(4)
EHW CI, linear		(-0.03, 0.16)	(-0.04, 0.09)	(-0.06, 0.04)
BME CI, linear		(-0.07, 0.20)	(-0.13, 0.17)	(-0.33, 0.31)
BME CI, quadratic		(-0.15, 0.37)	(-0.11, 0.28)	(-0.22, 0.30)
BSD CI, $K = 0.02$		(-0.06, 0.19)		
BSD CI, $K = 0.04$	(-0.08, 0.24)			
DD CI	(-0.05, 0.22)	(-0.06, 0.22)	(-0.13, 0.22)	(-0.32, 0.38)
DD p -value	0.350	0.350	1.000	1.000
Bandwidth h	2	3	6	∞

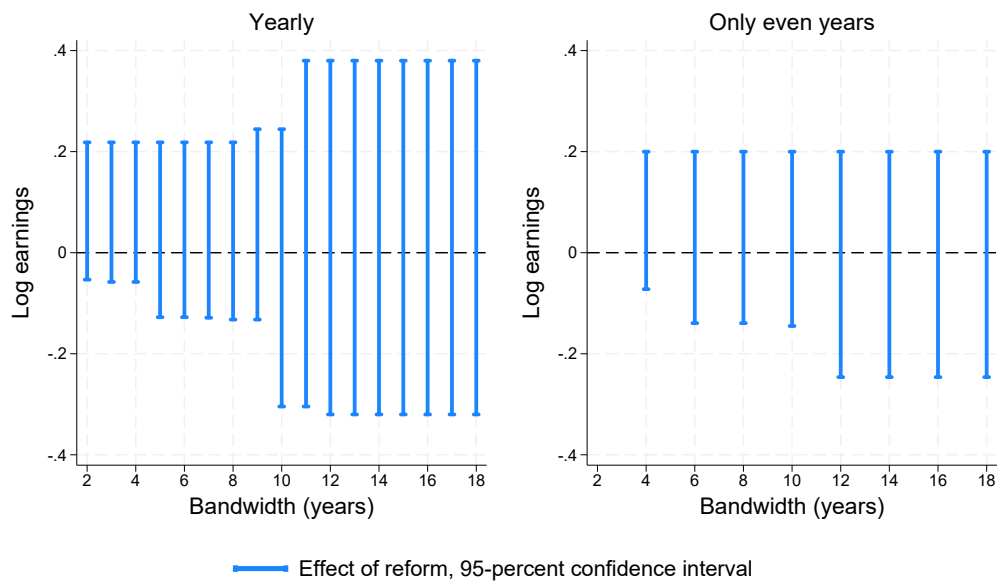
Notes: Based on data from Oreopoulos (2008). All results except those pertaining to our DD estimator are replications of figures from Table 3 in Kolesár and Rothe (2018).

Table 4: Effect of the 1947 UK school-leaving age reform on log earnings

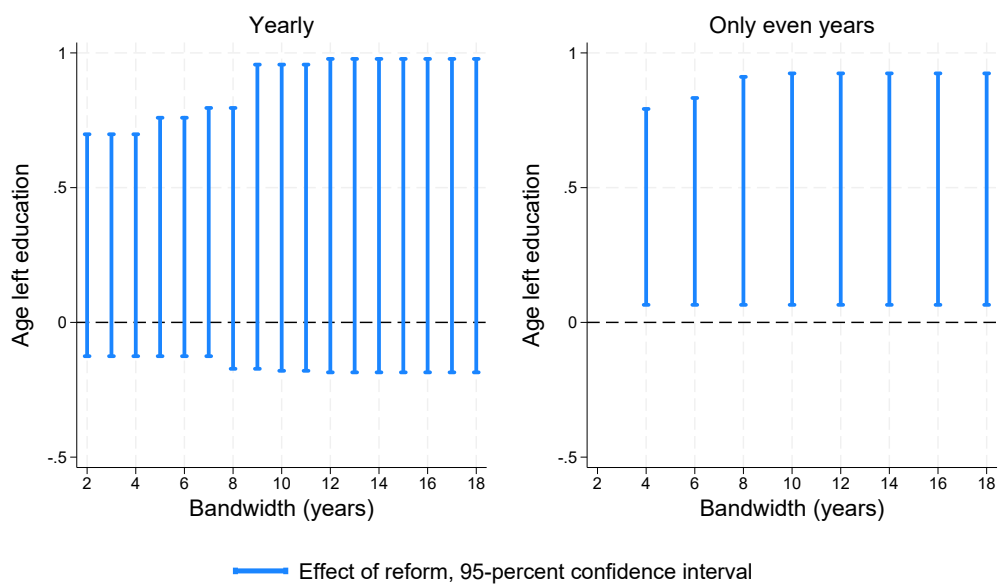
Table 4 shows that EHW confidence intervals are the narrowest at all bandwidths, and get narrower the larger the bandwidth. BME and DD confidence intervals are substantially wider, and tend to get wider the larger the bandwidth, as expected. DD produces somewhat wider or substantially narrower intervals than BME, depending on specification. The strong dependence of BME results on polynomial order indicates uncertainty about the appropriate functional form, which in turn makes it difficult to evaluate the identification assumption underlying BME. Finally, DD confidence intervals are slightly narrower than BSD for the more conservative choice of bound $K = 0.04$, but slightly wider for the less conservative $K = 0.02$.

The left panel of Figure 6a displays our DD confidence intervals for the full set of bandwidths, again clearly showing how width is weakly increasing in bandwidth. An important question is whether there is strong evidence that the reform affected completed education. Wide confidence intervals for the effect of the reform on earnings would not be surprising if there was no such evidence. Indeed, the left panel of Figure 6b shows that we cannot rule out at the 95 percent level that the reform did not affect years of schooling. Recall, however, that the reform took place in April 1947, meaning that some individuals who turned 14 in 1947 were still able to leave school in that year. Such a gradual phase-in of a reform will reduce statistical power.

To address the issue of gradual phase-in, one may instead restrict the data to



(a) Effect on earnings



(b) Effect on completed education

Figure 6: Effects of the school leaving age reform: Varying the bandwidth

Notes: The left panels vary the bandwidth using the original data from every year. The right panels do the same for a subset of the data that contains even years only.

even years, meaning that the effect of the reform is evaluated by comparing changes from 1946–1948 to changes between any other adjacent even years. Doing so results in stronger evidence that the reform did in fact cause an increase in completed education, as seen in the right panel of Figure 6b. There is still no strong evidence that the reform affected log earnings (right panel of Figure 6a). See also Figures A5–A6 for a display of raw data and double differences, both for earnings and completed education.

6 Conclusion

Regression discontinuity designs with discrete running variable are ubiquitous. Examples of such discrete running variables include raw points on a test, rank, number of employees, or year of birth. There currently appears to be no consensus among practitioners about how to perform inference in such settings, with available solutions relying on assumptions that are either very strong or difficult to evaluate. In this paper we have presented a new approach to inference for RD designs with a discrete running variable. The approach is based on an intuitive, general, and weak identification assumption and is easily implemented. Across a range of simulations, statistical power compares favorably with existing approaches.

Our approach highlights that evidence of a treatment effect in a discrete RD context requires unique behavior at the cutoff, especially with few support points. If there is a larger jump (of the same sign) away from the cutoff, then our method yields a p -value of unity. If researchers nevertheless argue in favor of a treatment effect, then the larger jump away from the cutoff will need to be explained.

We expect our double-differences approach to be useful beyond the setting studied in this paper. For instance, the approach could be applied to continuous RD settings where observations are sparsely distributed and clustered along the running variable, such as in a spatial RD context where populations are concentrated in cities. The approach is also applicable to event studies and difference-in-differences, where it enables statistical inference without the need to impose a parallel trends assumption.

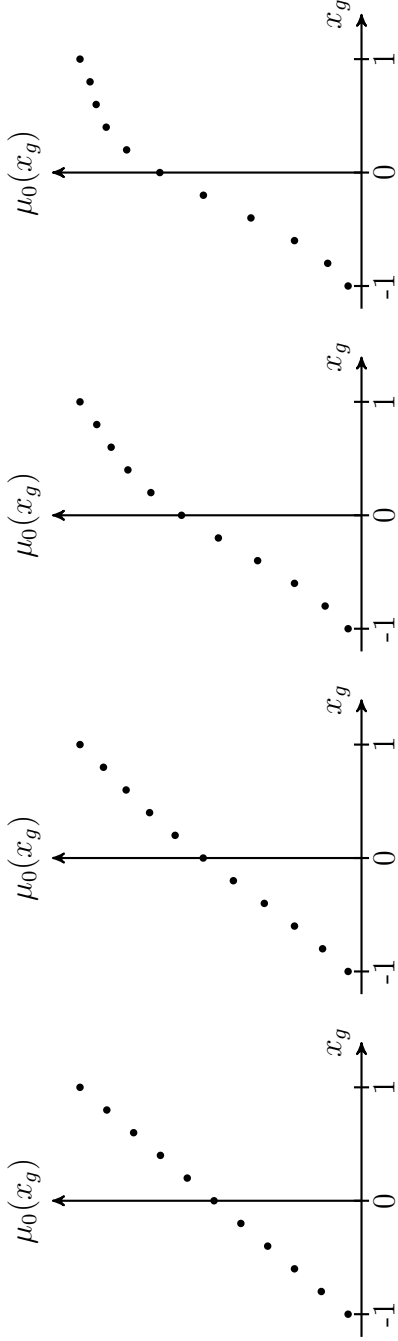
References

- Armstrong, T. B. and Kolesár, M. (2020). Simple and honest confidence intervals in nonparametric regression. *Quantitative Economics*, 11(1):1–39.
- Cao, Y., Fischer-Weckemann, B., Geyer, J., and Ziebarth, N. (2026). Fundamentally Reforming the DI System: Evidence from Germany. *The Review of Economics and Statistics*, pages 1–45.

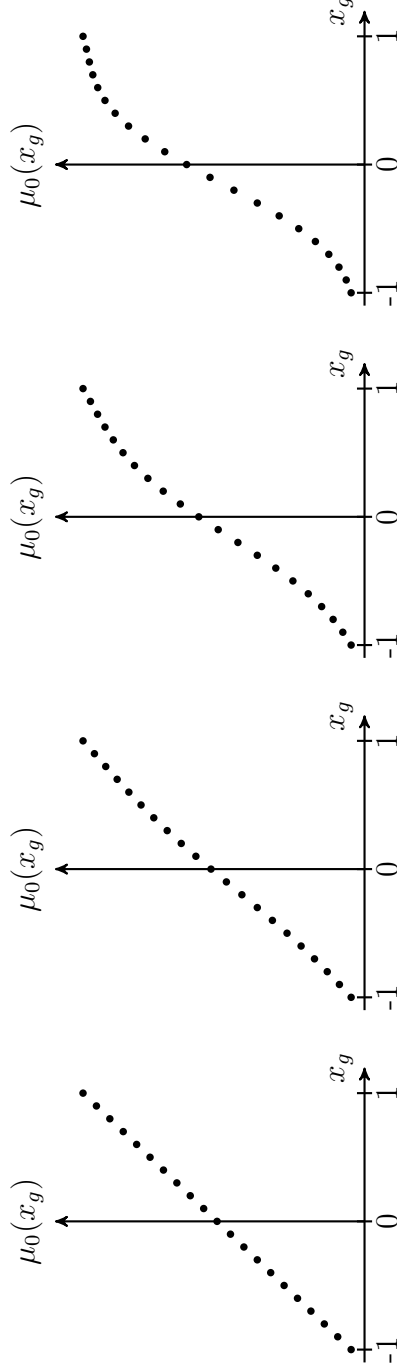
- Casella, G. and Berger, R. L. (2002). *Statistical Inference*. Duxbury, second edition.
- Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2020). *A Practical Introduction to Regression Discontinuity Designs: Foundations*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.
- Cattaneo, M. D., Idrobo, N., and Titiunik, R. (2024). *A Practical Introduction to Regression Discontinuity Designs: Extensions*. Elements in Quantitative and Computational Methods for the Social Sciences. Cambridge University Press.
- Cattaneo, M. D. and Titiunik, R. (2022). Regression Discontinuity Designs. *Annual Review of Economics*, 14(1):821–851.
- Eckles, D., Ignatiadis, N., Wager, S., and Wu, H. (2025). Noise-induced randomization in regression discontinuity designs. *Biometrika*, 112(2):asaf003.
- Figlio, D. and Ozek, U. (2024). The Unintended Consequences of Test-Based Remediation. *American Economic Journal: Applied Economics*, 16(1):60–89.
- Ganong, P. and Jäger, S. (2018). A Permutation Test for the Regression Kink Design. *Journal of the American Statistical Association*, 113(522):494–504.
- Graetz, G., Öckert, B., and Skans, O. N. (2025). Parental Education and the Responses to Higher SAT Scores. *Journal of Human Resources*, 60(6):1971–2007.
- Howell, S. T., Rathje, J., Reenen, J. V., and Wong, J. (2025). Opening Up Military Innovation: Causal Effects of Reforms to US Defense Research. *Journal of Political Economy*, 133(11):3605–3651.
- Kamal, F., Sundaram, A., and Tello-Trillo, C. J. (2026). Family-Leave Mandates and Female Labor at U.S. Firms: Evidence from a Trade Shock. *The Review of Economics and Statistics*, pages 1–17.
- 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., Sánchez, F., and Álvarez Arango, L. E. (2026). Financial Aid and Upward Mobility: Evidence from Colombia’s Ser Pilo Paga. *Journal of Political Economy*, 134(7):2119–2165.
- Manski, C. F. and Pepper, J. V. (2018). How Do Right-to-Carry Laws Affect Crime Rates? Coping with Ambiguity Using Bounded-Variation Assumptions. *The Review of Economics and Statistics*, 100(2):232–244.

- Montiel Olea, J. L. and Plagborg-Møller, M. (2019). Simultaneous confidence bands: Theory, implementation, and an application to SVARs. *Journal of Applied Econometrics*, 34(1):1–17.
- Mountjoy, J. (2026). Marginal Returns to Public Universities. *The Quarterly Journal of Economics*, 141(1):429–497.
- Oreopoulos, P. (2006). Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter. *American Economic Review*, 96(1):152–175.
- Oreopoulos, P. (2008). Estimating Average and Local Average Treatment Effects of Education when Compulsory Schooling Laws Really Matter: Corrigendum. *American Economic Review*.
- Rabaté, S., Jongen, E., and Atav, T. (2024). Increasing the Retirement Age: Policy Effects and Underlying Mechanisms. *American Economic Journal: Economic Policy*, 16(1):259–91.
- Rambachan, A. and Roth, J. (2023). A More Credible Approach to Parallel Trends. *The Review of Economic Studies*, 90(5):2555–2591.
- Sejas-Portillo, R., Moro, M., and Stowasser, T. (2025). The Simpler the Better? Threshold Effects of Energy Labels on Property Prices and Energy Efficiency Investments. *American Economic Journal: Applied Economics*, 17(1):41–89.

Appendix



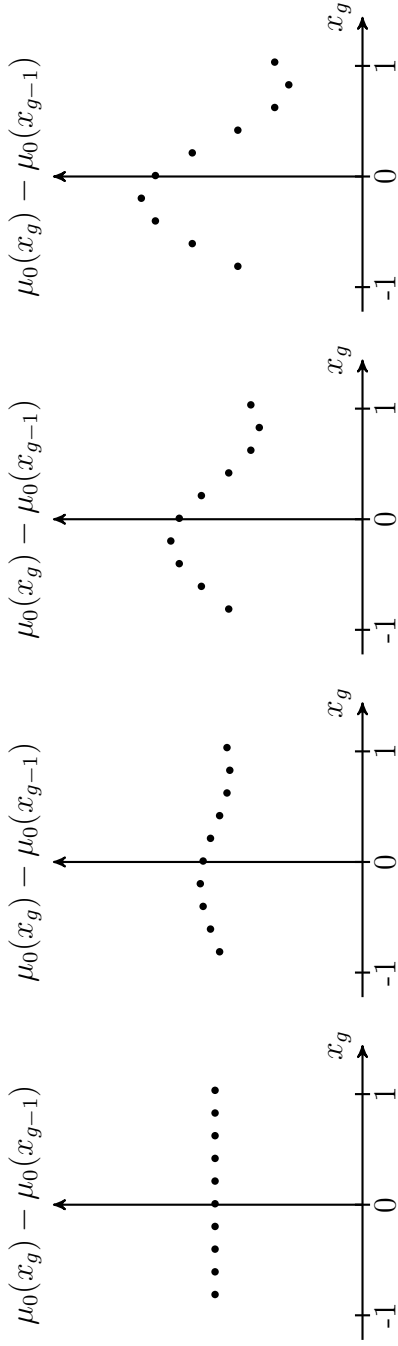
(a) Number of support points: $G = 11$



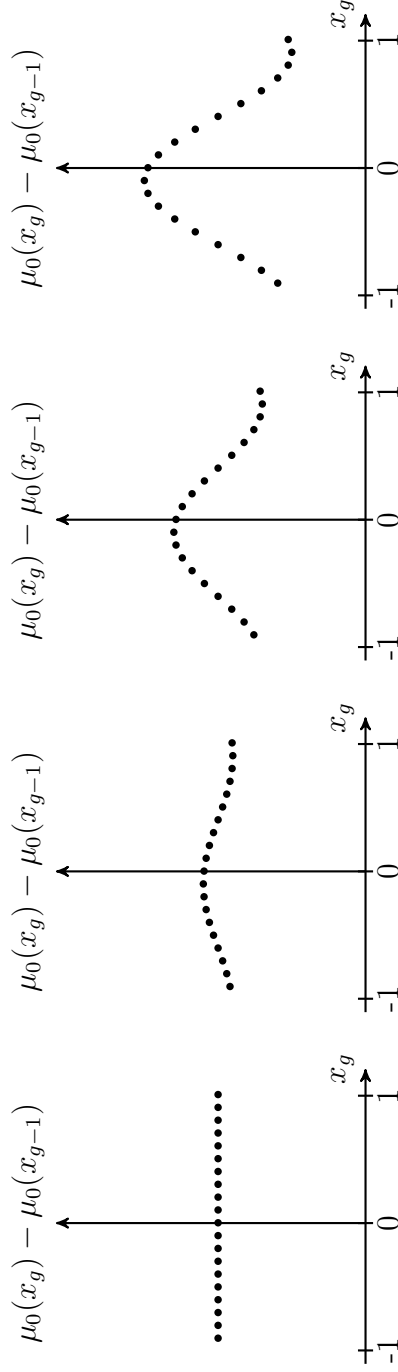
(b) Number of support points: $G = 21$

Figure A1: CEFs from the data generating process in equation (36)

Notes: From left to right, the parameter values are $\lambda \in \{0, 0.05, 0.15, 0.25\}$. CEFs are shown for the case when $j = 1$.



(a) Number of support points: $G = 11$



(b) Number of support points: $G = 21$

Figure A2: First differences of the data generating process in equation (36)

Notes: From left to right, the parameter values are $\lambda \in \{0, 0.05, 0.15, 0.25\}$. CEFs are shown for the case when $j = 1$.

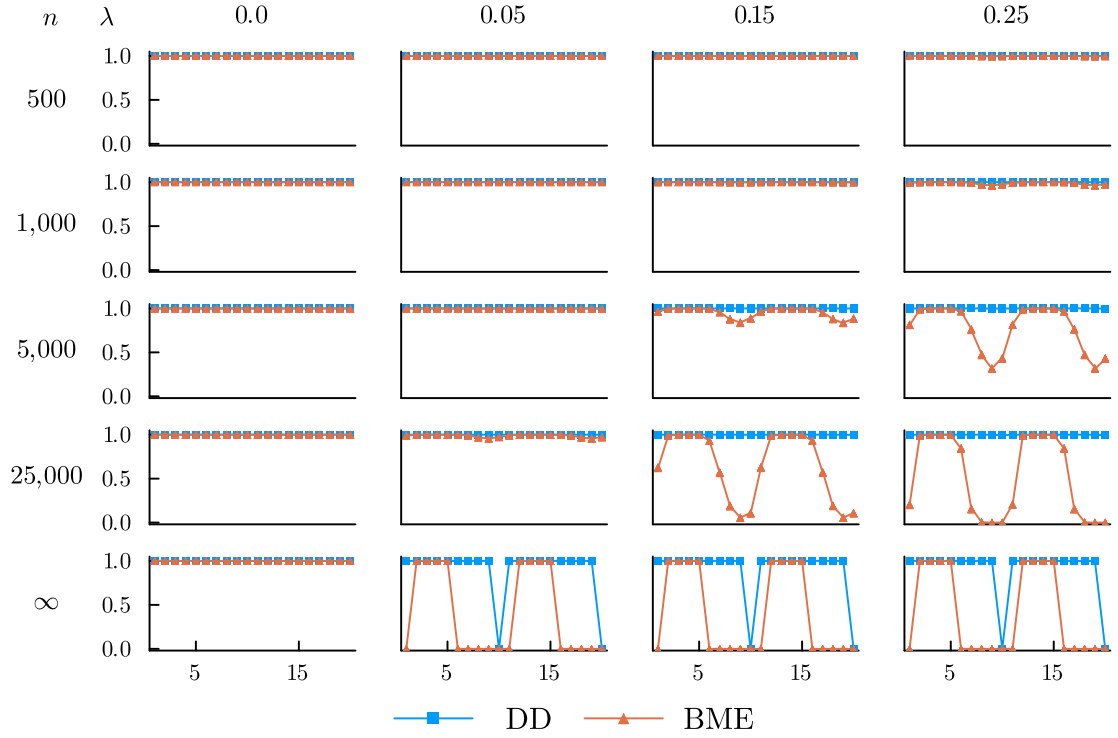
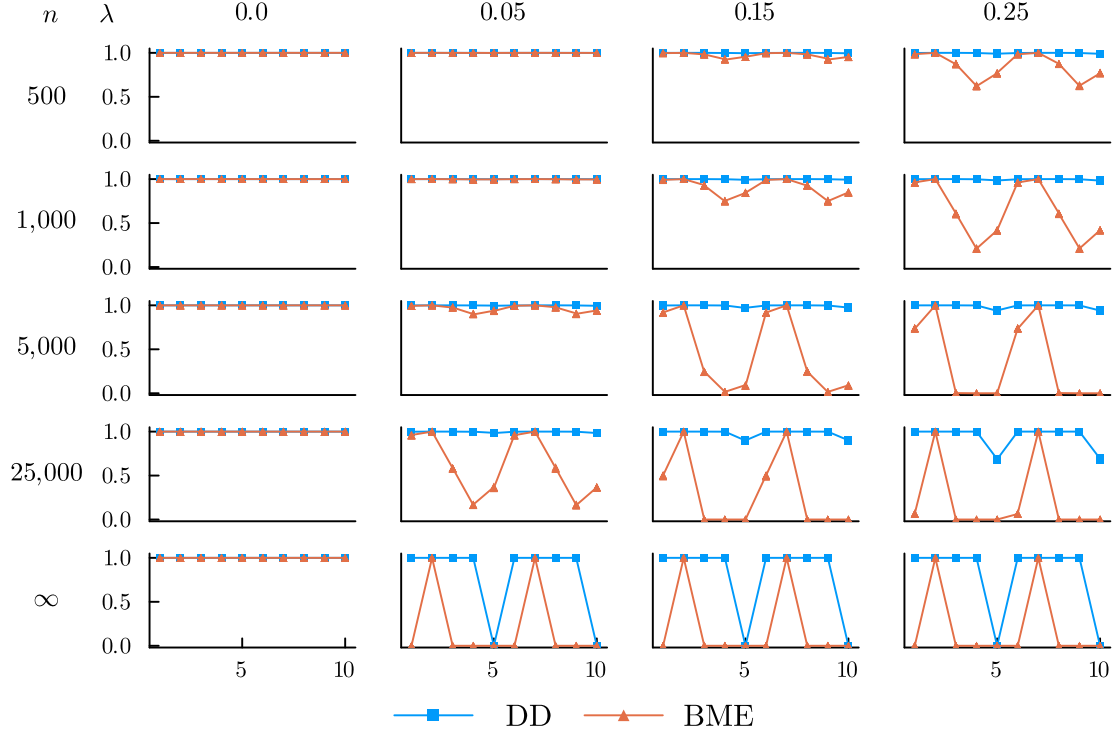


Figure A3: Coverage of 95-percent confidence intervals against j

Notes: Within each panel, coverage rates as a function of j (causing phase shifts via ζ_j) are shown for values of λ (left to right) and n (top to bottom).

G	$\lambda = 0.0$						$\lambda = 0.05$					
	DD (1)	BME (2)	BSD (3)	BSD* (4)	LL (5)	LL Δ (6)	DD (7)	BME (8)	BSD (9)	BSD* (10)	LL (11)	LL Δ (12)
Panel A: Normalized standard error, averaged over values of the phase shifter, ζ_j												
11	500	.154	.143	.060	1.864	.078	.158	.146	.102	1.869	.060	.078
11	1,000	.108	.100	.042	1.314	.055	.113	.104	.088	1.313	.042	.055
11	5,000	.048	.044	.019	.584	.025	.056	.051	.068	.590	.019	.025
11	25,000	.022	.020	.008	.261	.011	.032	.030	.059	.278	.008	.011
11	∞	.000	.000	.000	.000	.000	.015	.016	.052	.059	.000	.000
21	500	.230	.223	.058	1.271	.072	.230	.225	.095	1.276	.058	.072
21	1,000	.160	.154	.041	.894	.051	.161	.157	.081	.898	.041	.051
21	5,000	.070	.068	.018	.398	.023	.072	.073	.062	.407	.018	.023
21	25,000	.031	.030	.008	.178	.010	.035	.040	.053	.196	.008	.010
21	∞	.000	.000	.000	.000	.000	.008	.022	.047	.052	.000	.000
Panel B: Coverage of 95% CI, averaged over values of the phase shifter, ζ_j												
11	500	1.000	1.000	.949	1.000	.945	1.000	.999	.990	1.000	.843	.935
11	1,000	1.000	1.000	.949	1.000	.947	1.000	.998	.995	1.000	.738	.926
11	5,000	1.000	1.000	.950	1.000	.950	.999	.963	1.000	1.000	.320	.839
11	25,000	1.000	1.000	.950	1.000	.950	.996	.615	1.000	1.000	.179	.495
11	∞	1.000	1.000	.950	1.000	.950	.800	.200	1.000	1.000	.000	.000
21	500	1.000	1.000	.948	1.000	.945	1.000	1.000	.989	1.000	.857	.938
21	1,000	1.000	1.000	.950	1.000	.948	1.000	1.000	.994	1.000	.765	.933
21	5,000	1.000	1.000	.950	1.000	.950	1.000	1.000	.999	1.000	.357	.875
21	25,000	1.000	1.000	.950	1.000	.950	1.000	.988	1.000	1.000	.148	.605
21	∞	1.000	1.000	.950	1.000	.950	.900	.400	1.000	1.000	.000	.000

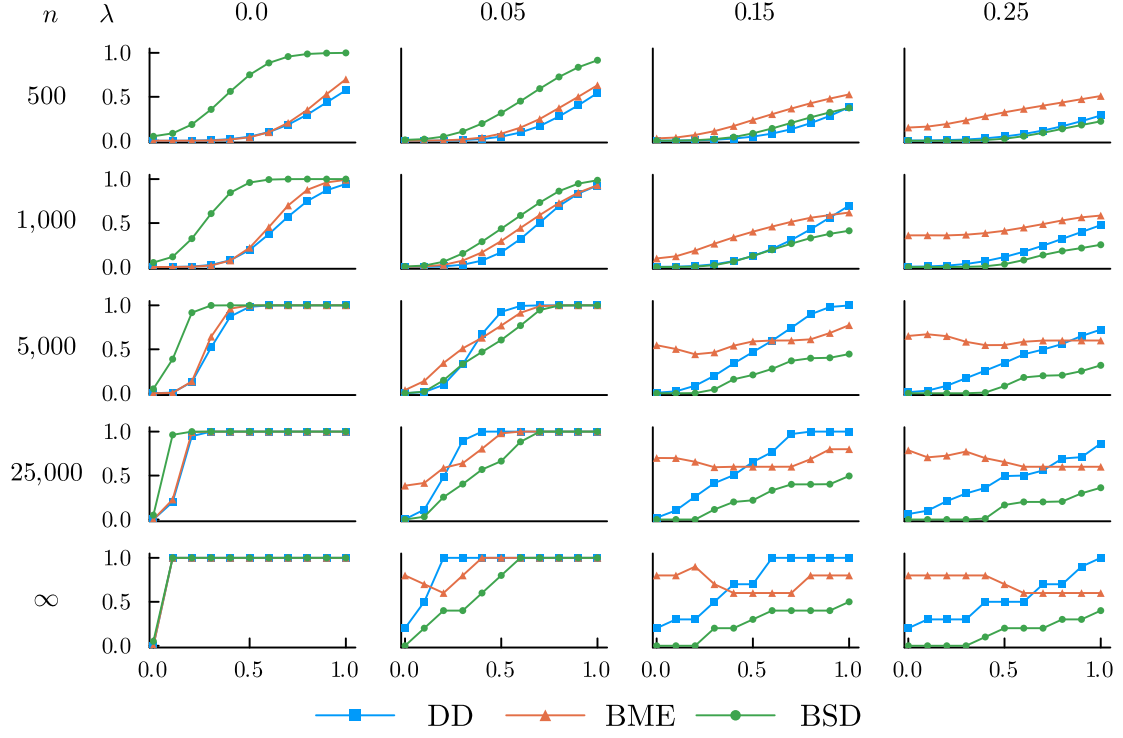
Notes: BSD* refers to the BSD method with data-driven choice of K as suggested by Armstrong and Kolesár (2020). LL is local linear regression with uniform kernel within bandwidth and robust standard errors. LL Δ is the same as LL, but where a triangular kernel has been used instead.

Table A1: Normalized standard errors and coverage of 95-percent confidence intervals

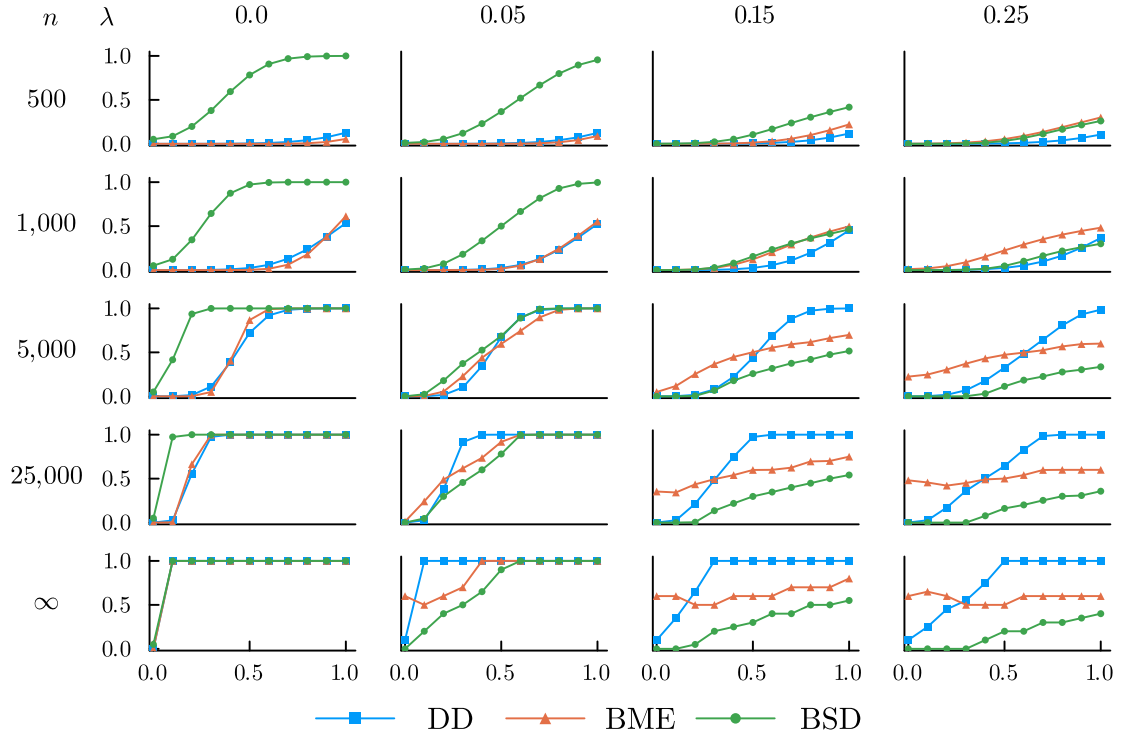
$\lambda = 0.15$													
G	$\lambda = 0.15$						$\lambda = 0.25$						
	DD (1)	BME (2)	BSD (3)	BSD* (4)	LL (5)	LL Δ (6)	DD (7)	BME (8)	BSD (9)	BSD* (10)	LL (11)	LL Δ (12)	
<i>Panel A: Normalized standard error, averaged over values of the phase shifter, ζ_j</i>													
11	500	.177	.164	.207	1.888	.060	.078	.202	.190	.311	1.926	.061	.078
11	1,000	.135	.125	.192	1.345	.043	.055	.162	.154	.296	1.395	.043	.055
11	5,000	.083	.079	.172	.647	.019	.025	.112	.111	.276	.721	.019	.025
11	25,000	.062	.061	.163	.355	.009	.011	.092	.094	.267	.446	.009	.011
11	∞	.046	.049	.156	.176	.000	.000	.077	.082	.260	.293	.000	.000
21	500	.235	.240	.188	1.301	.058	.072	.244	.265	.282	1.343	.059	.072
21	1,000	.167	.176	.174	.929	.041	.051	.178	.205	.267	.983	.042	.051
21	5,000	.084	.103	.155	.464	.018	.023	.098	.142	.248	.539	.019	.023
21	25,000	.049	.079	.147	.272	.008	.010	.064	.122	.240	.360	.008	.010
21	∞	.024	.065	.140	.156	.000	.000	.040	.108	.233	.261	.000	.000
<i>Panel B: Coverage of 95% CI, averaged over values of the phase shifter, ζ_j</i>													
11	500	.999	.973	.999	1.000	.349	.848	.997	.850	1.000	1.000	.201	.680
11	1,000	.998	.902	1.000	1.000	.227	.751	.996	.640	1.000	1.000	.179	.495
11	5,000	.994	.453	1.000	1.000	.169	.344	.988	.348	1.000	1.000	.133	.190
11	25,000	.980	.300	1.000	1.000	.094	.136	.938	.213	1.000	1.000	.015	.052
11	∞	.800	.200	1.000	1.000	.000	.000	.800	.200	1.000	1.000	.000	.000
21	500	1.000	1.000	.999	1.000	.386	.876	1.000	.999	1.000	1.000	.212	.759
21	1,000	1.000	.999	1.000	1.000	.252	.812	1.000	.989	1.000	1.000	.149	.603
21	5,000	1.000	.951	1.000	1.000	.109	.437	1.000	.776	1.000	1.000	.048	.235
21	25,000	1.000	.647	1.000	1.000	.021	.173	.999	.519	1.000	1.000	.000	.101
21	∞	.900	.400	1.000	1.000	.000	.000	.900	.400	1.000	1.000	.000	.000

Notes: BSD* refers to the BSD method with data-driven choice of K as suggested by Armstrong and Kolesár (2020). LL is local linear regression with uniform kernel within bandwidth and robust standard errors. LL Δ is the same as LL, but where a triangular kernel has been used instead.

Table A2: Normalized standard errors and coverage of 95-percent confidence intervals



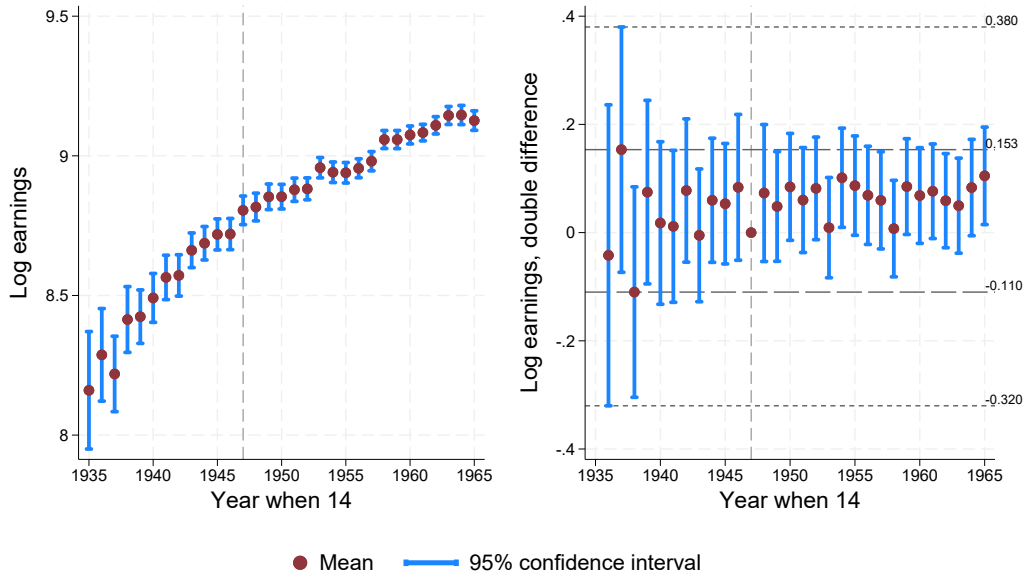
(a) Number of support points: $G = 11$



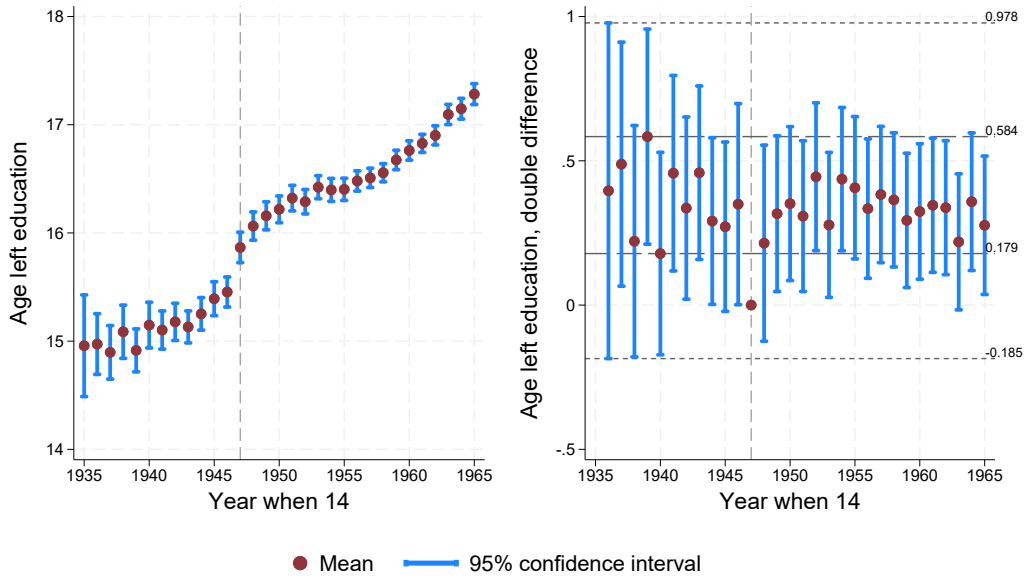
(b) Number of support points: $G = 21$

Figure A4: Rejection rate by normalized treatment effect

Notes: Within each panel, rejection rates averaged across phase shifts as a function of normalized treatment effects, $\tau/\sqrt{\mathbb{V}(\varepsilon)}$, are shown for values of λ (left to right) and n (top to bottom).



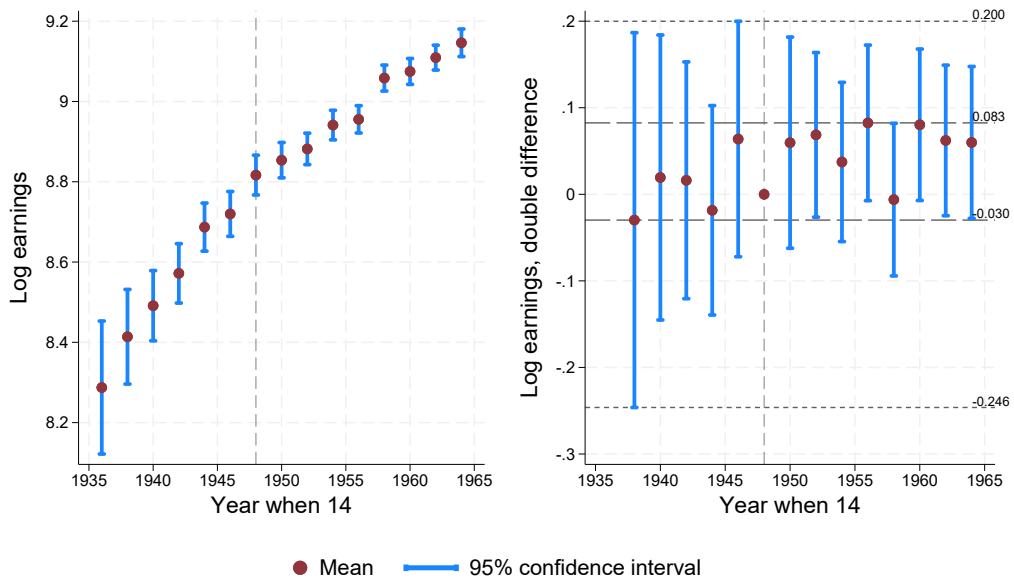
(a) Effect on earnings



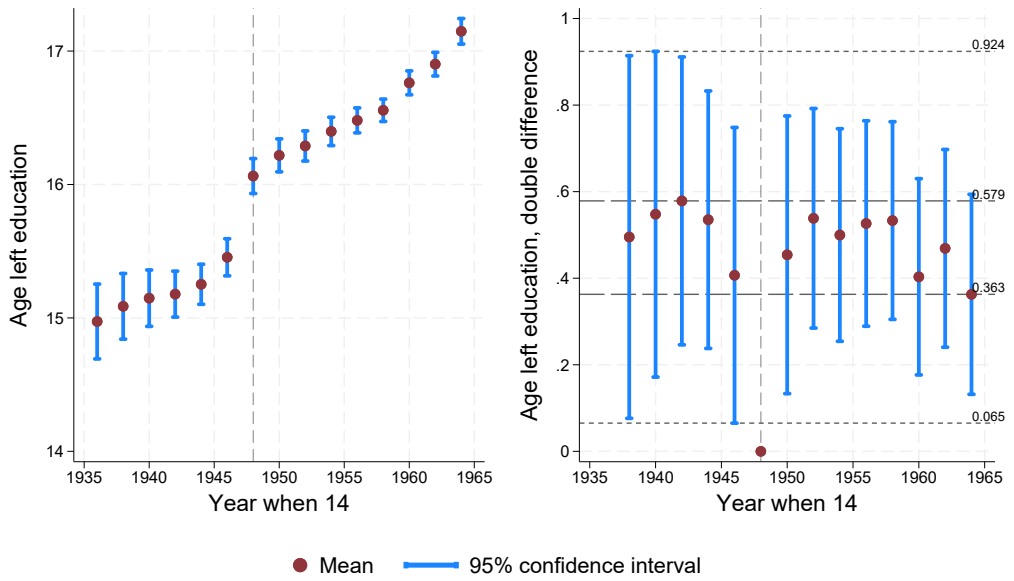
(b) Effect on completed education

Figure A5: School leaving age reform, earnings and completed education

Notes: In the left panels, log earnings and school leaving age are plotted against the year individuals turned 14. In April 1947, the UK raised the minimum school leaving age from 14 to 15. Data are from Oreopoulos (2008). In the right panels, double differences are plotted. Short-dashed lines mark 95-percent confidence intervals for the treatment effect. Long-dashed lines mark bounds on the treatment effect.



(a) Effect on earnings



(b) Effect on completed education

Figure A6: School leaving age reform, earnings and completed education, even years

Notes: In the left panels, log earnings and school leaving age are plotted against the year individuals turned 14. In April 1947, the UK raised the minimum school leaving age from 14 to 15. Data are from Oreopoulos (2008) but using only even years. In the right panels, double differences are plotted. Short-dashed lines mark 95-percent confidence intervals for the treatment effect. Long-dashed lines mark bounds on the treatment effect.