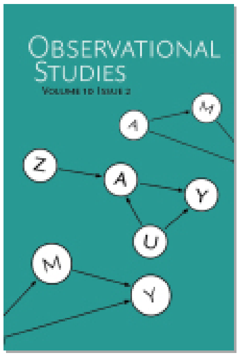
Constructing multiple, independent analyses in the regression discontinuity design with multiple cutoffs



Youjin Lee, Chichun Tan, Bikram Karmakar

Observational Studies, Volume 10, Issue 2, 2024, pp. 63-91 (Article)

Published by University of Pennsylvania Press
DOI: https://doi.org/10.1353/obs.2024.a946585

Observational Studies 10.2 (2024): 63-92 Submitted Oct 22 2024; Published Dec 2024

Constructing multiple, independent analyses in the regression discontinuity design with multiple cutoffs

Youjin Lee youjin_lee@brown.edu
Department of Biostatistics
Brown University
Providence, RI 02912, USA

Chichun Tan chichun_tan@brown.edu
Department of Biostatistics
Brown University
Providence, RI 02912, USA

Bikram Karmakar bkarmakar@ufl.edu
Department of Statistics
University of Florida
Gainesville, FL 32611, USA

Abstract

The regression discontinuity (RD) design is a commonly used non-experimental approach for evaluating policy or program effects. However, this approach heavily relies on the untestable assumption that distribution of confounders or average potential outcomes near or at the cutoff are comparable. When there are multiple cutoffs that create several discontinuities in the treatment assignments, factors that can lead this assumption to the failure at one cutoff may overlap with those at other cutoffs, invalidating the causal effects from each cutoff. In this study, we propose a novel approach for testing the causal hypothesis of no RD treatment effect that can remain valid even when the assumption commonly considered in the RD design does not hold. We propose leveraging variations in multiple available cutoffs and constructing a set of instrumental variables (IVs). We then combine the evidence from multiple IVs with a direct comparison under the local randomization framework. This reinforced design that combines multiple factors from a single data can produce several, nearly independent inferential results that depend on very different assumptions with each other. Our proposed approach can be extended to a fuzzy RD design. We apply our method to evaluate the effect of having access to higher achievement schools on students' academic performances in Romania.

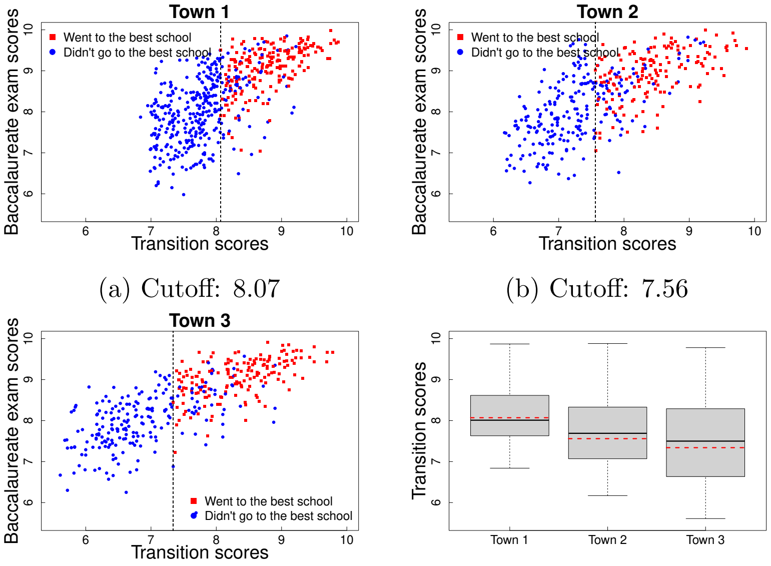
Keywords: Evidence factors, Local randomization, Regression discontinuity, Replication

1. Motivating example: the effect of attending higher achievement schools on students' performance

In education and social sciences, it is often of interest whether access to higher-achieving schools has a causal effect on students' future academic performance (Dale and Krueger, 2002; Cullen et al., 2005; Clark, 2010; Dobbie and Fryer Jr, 2014). Existence of such effect could also have economic impacts by influencing housing prices and demographic compositions at school district borders (Abdulkadiroğlu et al., 2014; Laliberté, 2021). The most ideal setting to evaluate this causal effect is to randomize students into a "better school" (e.g., Duflo et al. (2011)) and compare academic performance between a group of students who went to a better school and did not go to a better school. With this design, concerns about confounding factors can be reduced, which could result in systematic differences between the two different groups of students not due to attending a better school. Alternatively, random assignment of encouragement to attend a better school, as in the Moving to Opportunity (MTO) study (Chetty et al., 2016), could be used to enhance children's neighborhood environments. However, such experiments are often infeasible or expensive in real world settings. It is challenging to identify the effect of going to a better school unless a very compelling study design is adopted so that students who went to a better school are comparable to those who did not go there otherwise in terms of confounders. Instead, researchers often rely on the designs in which school assignments are affected by students' selections and school admission process. Fortunately, one can leverage natural experiments embedded in the admission process in which admission to prestigious schools is dominantly determined by students' scores and schools' own cutoffs. This process induces a discontinuity in the treatment assignment with respect to students' scores and creates a regression discontinuity (RD) design (Thistlethwaite and Campbell, 1960; Hahn et al., 2001).

In the sharp RD design, each unit's treatment assignment depends solely on the values of their running variable – units with values for the running variable (e.g., students' scores) above the cutoff receive the treatment while units with the values below the cutoff receive the control (Thistlethwaite and Campbell, 1960; Hahn et al., 2001). One can exploit such discontinuity in the treatment assignment (i.e., going to a better school or not) and then compare the future outcomes of students who were barely admitted to a better school to those who barely missed admission. As students' preferences, as well as their scores, also affect their school choices, school selection processes are often fuzzy RD designs where having a running variable value above the cutoff would increase the probability of receiving the treatment but does not solely determine the treatment assignment (Hahn et al., 2001). To illustrate one example, Lucas and Mbiti (2014) used students' scores on the national standardized primary school exit exam as a running variable that determined the eligibility for admission to the most prestigious secondary schools in Kenya. As cutoffs vary by schools and school districts, the study used the standardized running variable by subtracting the cutoff value, and then obtained the causal effect estimates through two-stage least squares models. This analysis is based on the assumption that the indicator of having scores above the cutoff serves as an instrumental variable (IV) (Angrist et al., 1996) for attending prestigious schools. Their findings suggested little evidence of the effects of attending the best Kenyan secondary schools in the districts on students' academic achievement.

Contributing to this line to investigation, we use Romanian administrative data on admission to study the effect of attending the best high school in a student's town on their performance on the baccalaureate exam. The baccalaureate exam, which is taken at the end of high school in July, is a critical criterion for university admissions. If attending prestigious schools is found to have a significant effect, government-level interventions in the education system, such as reallocating educational resources and incentivizing teachers, could be considered to improve school performance and reduce disparities between schools. Moreover, it could provide motivation to reevaluate the admission guidelines for those schools



(a) Cutoff: 8.07

(b) Cutoff: 7.56

(c) Cutoff: 7.34

(d)

Figure 1: (a)-(c): Distributions of transition scores (the running variable) and baccalaureate exam scores (the outcome variable) of Romanian students from three towns with 539, 381, and 371 students. Square dots denote students who went to the best school in their town and round dots denote students who did not go to the best school. In each figure, a dashed vertical line indicates the cutoff (the minimum transition score required to be admitted to the best school). (d): The distribution of transition scores and the cutoff (dashed horizontal lines) of three towns.

so that a larger population can benefit from it. Here the running variable is each student's score used for high school admission, so-called "transition score", which is the average of their performance on a national exam taken in grade 8 and their grade point average from grades 5-8 (Pop-Eleches and Urquiola, 2013). Then the cutoff for entering each school is determined by the minimum transition score among the students who are assigned to that school at the same admission period. The best high school is identified by the school in the town that has the highest minimum transition score required for admission. In this context, the cutoff can be viewed as a random variable, whose value is influenced by various factors, including student's performance and the size of schools in their town. If a student's score is higher than the cutoff of their town, then she could enter a better school while she could not enter the school if the score is lower than the cutoff. The administrative data provide students' transition scores and the high school they attended. We pull the data from 13 towns in Romania. Their cutoffs vary across towns but the actual treatment assignment (i.e., going to a better school or not) is student-level.

Figures 1(a)-(c) illustrate the distribution of students' transition scores and their baccalaureate exam scores from three randomly chosen towns in the dataset. We observe that not all students whose transition scores exceeded the cutoff went to the best school; some students chose to go to other schools even though their scores were higher than the minimum scores required. This could be due to a lack of information on school quality, and proximity, and household's economic status (Ainsworth et al., 2023). It is also likely that students were unable to attend their most preferred school among the eligible schools as their choice has been full. In such a case, they were assigned to their next preferred school. However, within these three towns, there was no student who went to the best school even if their scores were below the cutoff (i.e., there are no square dots in the left side of the vertical line). This presents a fuzzy RD design under non-compliance. Table 1 presents the number of students who were eligible for attending the best school and who actually went to the school.

	Town 1	Town 2	Town 3
The number of students	539	381	371
Cutoff	8.07	7.56	7.34
The number (percentage) of students with transition scores greater than the cutoff	255 (47.3%)	213 (55.9%)	207 (55.8%)
The number (percentage) of students who went to a better school	195 (36.2%)	154 (40.4%)	162 (43.7%)

Table 1: Statistics of high school admissions in three towns in Romania.

As in the study of Lucas and Mbiti (2014), we can analyze the RD effect as if there were a single cutoff value (e.g., zero) and use the standardized transition score as a new running variable. Incorporating the data from different towns in this way could enhance statistical power and generalizability of the findings compared to the study relying on a single town. However, this standardization of the transition score discards any information on variations in cutoffs assignments across different towns by pretending that they were all zeros. The boxplot in Figure 1(d) and the third row of Table 1 suggest that comparatively large number of students from towns with lower cutoffs had their scores above the cutoff, even though transition scores and cutoffs are seemingly positively correlated. Consequently, pooling data from multiple towns may result in a disproportionate distribution of treated and control students across the towns; for example, more number of treated students would be pulled from towns with lower scores. Moreover, when using standardized scores, students from Town 1 with scores just above the new cutoff (i.e., zero) could have very different values of transition scores compared to those from Town 3 with scores just below the new cutoff. These all could lead to a bias if town-specific factors confound the causal effect.

In our work, in addition to using the actual treatment assignment, we use the cutoff assignment as a potential IV that affected the admission to a better school but would not necessarily have a direct effect on students' future performances.

2. Evidence factors analysis in regression discontinuity design

2.1 Regression discontinuity design and two different frameworks

The RD design has been widely used for evaluating causal effects of program or policy interventions in public health (e.g., Venkataramani et al. (2016)), social sciences (e.g., Lee and Lemieux (2014); Gerber and Hopkins (2011)), and education (e.g., Moss and Yeaton (2006); Robinson (2011); Figlio et al. (2018); Díaz and Zubizarreta (2023)). There are two frameworks used to analyze the RD design. The first is the continuity-based design (Hahn et al., 2001). Their key underlying assumption is that the average potential outcomes of treated and control units *at the cutoff* are the same if their treatment assignment would not change. Under this continuity assumption, the most popular inferential approach is via local polynomial non-parametric regressions (Fan and Gijbels, 1996) that fit the outcome on a multiple-order polynomial expansion of a running variable, each for treated and control units near the cutoff. Then the difference in the estimated intercept coefficients in the two polynomial regressions can be interpreted as the RD treatment effect at the cutoff.

The second framework is the local randomization design (Cattaneo et al., 2015). This framework assumes that there exists a *window* around the cutoff within which units are nearly randomly assigned to treatment. In this way, the RD design can be viewed as a randomized experiment, also called a *tie-breaking* experiment (Trochim, 1984), in which units below and above the cutoffs are randomly different (after some adjustment using covariates) as long as they are within the window. Once the appropriate window is chosen, several methods for inference can be applied based on the assumptions and the type of causal hypotheses. For example, suppose that units within the window are matched on all available confounders, and the sharp null hypothesis is of interest. Then randomization-based inference on matched pairs can be adopted to calculate exact p-values, which utilizes the known distribution of the random assignment of treatments within matched pairs. As opposed to continuity-based approach, the local randomization does not necessarily require the specifications of regressions. Moreover, inferential results derived from this framework can be easily generalizable to other target populations as long as units in the window are only randomly different other than the treatment assignment (Díaz and Zubizarreta, 2023). Our work focuses on the local randomization framework.

2.2 Multiple cutoffs, bias, and instrumental variables

Across different disciplines, including our motivating example of school admission, RD designs often involve different cutoffs among units that determine or affect each unit's treatment assignment. For example, while studying the effect of winning an election, RD design can be used with vote shares as a running variable. However, the minimum vote share required for candidates to win an election may differ by election districts (Gerber and Hopkins, 2011; Cattaneo et al., 2016b). Similarly, in education, admission to specific schools or programs is frequently determined by students' academic or poverty scores, and these cutoffs can vary across different levels of educational programs or across schools (Van der Klaauw, 2002). Even though incorporating different populations with different cutoffs may introduce heterogeneity in causal effects of interest, RD designs with multiple cutoffs typically enhance statistical power compared to the analysis using a single cutoff (e.g., that is limited to a single district).

There are several different approaches commonly used for multiple cutoffs (hereafter multi-cutoff) setting. One can analyze the RD treatment effect separately for subsets of units with the same cutoff and then combine the results together from multiple subgroups. This approach allows us to examine the causal effect multiple times with multiple, independent sets of observations. Under the local randomization-based framework, multi-cutoff RD design can be regarded as performing multiple randomized experiments at each cutoff.

Rather than conducting several analyses, it is also a common practice to analyze the multi-cutoff RD design as if all the units had a single cutoff, e.g., zero, and use a standardized running variable, e.g., the original value of a running variable minus a cutoff (Pop-Eleches and Urquiola, 2013; Önder and Shamsuddin, 2019; Melguizo et al., 2016; Cattaneo et al., 2016b). This approach was adopted in the aforementioned study of Lucas and Mbiti (2014). Despite its convenience and potential gains in sample sizes, if there exists a bias that invalidates the analysis at one cutoff, the whole, combined analysis would also be biased. For instance, in the study of Lucas and Mbiti (2014), suppose that a grader from *one* school district could manipulate some students' scores based on subjective judgment to influence their admission into prestigious secondary schools, or that the cutoff was adjusted to admit specific students within that district. Then even if students from other school districts were randomly different around their own cutoffs, without knowing which district had such manipulation, the combined analysis would undermine the key RD assumptions.

It is commonly suspected that units with a running variable just below and above the cutoff are systematically different not necessarily due to the treatment assignment but potentially due to unobserved factors. Consequently, when using local randomization, it is not guaranteed that the treated and control units are just randomly different. This could easily introduce bias in the estimated causal effect. To our knowledge, there is no RD approach available in the literature that remains valid when this assumption breaks. In this work, we propose using IVs to separate the RD causal effect from this common type of bias. The use of IV approaches in the RD setting has been dominantly discussed in a fuzzy RD design. In our proposed approach, we suggest different types of IVs that are not affected by the violation of the common assumption considered in the conventional, sharp or fuzzy RD analysis. We illustrate how multi-cutoff settings can provide multiple potential IVs, each of which is related to the treatment assignment but is believed not to have a direct effect on the outcome (i.e., cutoff exclusion restriction) under certain conditions. Even in the case where each of the potential IVs is violating the cutoff exclusion restriction, we demonstrate that bias from such invalid IVs would be very different from bias that is typically concerned in either the continuity-based or local randomization designs. We then propose combining multiple IVs with a direct comparison between the two treatment arms under the local randomization design. This approach can provide multiple pieces of evidence for testing the RD treatment effect, which could enhance robustness and testing power together.

In this work, we focus on testing the causal hypothesis of no RD treatment effect, which may be inverted to estimating the RD treatment effect. We specifically consider the sharp null hypothesis of no treatment effect for all units and incorporate the local randomization framework to our proposed approach. The sharp null hypothesis can be robust to a small number of units with a value of the running variable being around the cutoffs (Cattaneo and Titiunik, 2022). Consideration of the sharp null hypothesis is also useful in our setting as we aggregate data from multiple subpopulations (e.g., multiple towns), each with distinct cutoffs and potentially different treatment effects.

2.3 Evidence factors design

Researchers may be tempted to analyze the same RD design data using *both* the continuity and local randomization frameworks. For example, they might fit local polynomial regression models under the continuity-based design and also apply a Wilcoxon rank-sum test within the selected window around the cutoff under the local randomization framework. The former analysis can supplement a small number of samples within the window with samples out of the window by relying on the model, while the latter analysis can be more robust to model specification than the former regression analysis. Thus we would expect that these two analyses provide complementary evidence. Consequently, if the estimated causal effects are significant with a small p-value in both frameworks, this could suggest stronger and more reliable evidence against the null hypothesis than relying on a single approach.

In the context of RD design, the idea of corroborating causal conclusions from *coupling* multiple analyses was first discussed by Donald Campbell in Trochim (1984).

> "*The statistical properties of estimates that we get out of coupling designs can be better than those we get out of single approach designs. To put this more concretely, even had a tie-breaking randomization been permitted, a supplementary regression-discontinuity analysis would also have been desirable. In addition, a tie-breaking randomization no matter how few its cases will always add inferential strength to a regression-discontinuity analysis.*"

Even though *coupling* the two frameworks together seems promising, it may not always *"add inferential strength."* If the two results are highly probabilistically correlated under the null hypothesis, obtaining such agreeable results from them is not equivalent to having complementary evidence. Thus, congruence of two analyses for the same causal hypothesis will be most strengthened when the two inferential procedures are statistically independent under the null. Furthermore, these two analyses might agree only because they are wrong together. It is common in observational studies that evidence obtained from one analysis is incorrect or invalid for some reasons that could also invalidate the evidence from other analyses. In such a case, there could be no clear benefit from collecting evidence from multiple analyses. For example, if an uncontrolled abrupt change occurs around the cutoff, rendering units just below and just above incomparable, both the continuity-based and local randomization designs could easily result in bias, potentially falsely rejecting the null hypothesis. This abrupt change can also lead to sorting (Lee, 2008) that invalidates inference based on the continuity framework.

Consequently, the benefits of "*coupling designs*" could be maximized if evidence from each analysis is independent and does not overlap with each other. This enables multiple pieces of evidence to be easily combined without relying on complex multiple testing procedures while the combined results are still robust to the existence of invalid pieces. When multiple analyses for the same causal hypothesis satisfies these conditions, they are called *evidence factors* (Rosenbaum, 2010). However, it is challenging to meet these properties when conducting multiple analyses using a single data set, e.g., because in RD analysis, both local randomization and continuity-based analysis use same units just below and above the cutoff. In this work, we will propose a new method for evidence factors analysis for multi-cutoff RD setting, which allows us to construct multiple pieces of evidence using a single dataset while satisfying these desirable conditions.

We say that multiple analyses construct evidence factors when (i) each analysis testing the same null hypothesis results in p-values that are nearly independent under the null and (ii) invalidity of one analysis does not necessarily lead to invalidity of another analysis (Rosenbaum, 2010, 2011, 2017). When two or more evidence factors provide supportive evidence, then evidence for a causal effect can be strengthened. This is because obtaining significant results (e.g., p-values less than the significance level) from multiple and independent sources is less likely than observing a significant result from just one source under the null; and bias that might invalidate one source of evidence does not necessarily undermine the supportive evidence from the other factors.

In this work, rather than combining inferences from the two frameworks used in RD designs, we use both potential IVs and the local randomization design to produce several evidence factors for testing the same sharp null hypothesis. This design is called *reinforced design* (Karmakar et al., 2021) in which several IVs are used to construct multiple evidence factors in addition to one direct comparison. Our proposed approach leverages the reinforced design and allows some proposed IVs to violate their key assumptions (e.g., exclusion restriction). In that way, we can *"add inferential strength"* to the RD analysis, which is often lacking in the RD design due to its heavy reliance on a subset of units near the cutoff and also due to the violation of either continuity or local randomization assumptions that may be present across different cutoffs.

3. Setting and Notation

Let W be a running variable and D be a treatment indicator. Consider the setting where each unit has its own cutoff $C \in \mathcal{C}$, treating C as a random variable. Suppose that we have a finite number of cutoffs in observed dataset and denote the number of distinct cutoffs as q ($q \geq 2$), i.e., $|\mathcal{C}| = q$. Let $\mathcal{C} = \{c_1, \ldots, c_q\}$, where $c_1 < c_2 < \ldots < c_q$. We first consider a sharp RD design where each unit receives a treatment if and only if $W \geq C$; in other words, $D = \mathbb{I}(W \geq C)$ with $\mathbb{I}(\cdot)$ denoting an indicator function. Let Y denote the outcome of interest.

3.1 Proposed instrumental variables

We can propose up-to $(q-1)$ IVs when we have q distinct cutoffs: $Z_k = \mathbb{I}(C \leq c_{q-k})$ for $k = 1, 2, \ldots q-1$, indicating having a cutoff no larger than c_{q-k}. Each of Z_k can be considered as a potential IV if units with the cutoff no larger than c_{q-k} would be more likely to be assigned to the treatment than those with the cutoff larger than c_{q-k} under certain conditions.

Definition 1 *For $k = 1, 2, \ldots q-1$, we call $Z_k = \mathbb{I}(C \leq c_{q-k})$ a potential IV when it satisfies*

$$E\{\mathbb{I}(W \geq C) \mid C \leq c_{q-k}\} \quad > \quad E\{\mathbb{I}(W \geq C) \mid C > c_{q-k}\}. \tag{1}$$

In other words, each potential IV, Z_k, is positively associated with the treatment assignment. This condition can hold when C is independent of W.

Lemma 2 *When W is independent of C and $F_W(c_{q-k+1}) > F_W(c_{q-k})$ with F_W denoting a density function of W, then Equation (1) holds for $k = 1, 2, \ldots, q-1$.*

On the other hand, if W is nearly centered around C (e.g., a higher value of C is associated with a higher value of W), then this condition is rarely satisfied unless there are observed covariates available which make C and W conditionally independent. Regardless of (conditional) independence between W and C, however, one can easily test (conditional) associations between each of Z_k and D.

Each of the proposed IVs violates the exclusion restriction and/or no unmeasured confounding assumption of IVs when a cutoff C has a direct effect on the outcome and/or there are unmeasured confounders between the cutoff and the outcome of interest. In practice, it is likely that C is correlated with the running variable W that is often associated with the outcome so that Z_k (the cutoff being no larger than c_{q-k}) violates the IV assumptions unless W is properly controlled for. Roughly speaking, the bias coming from the violation of the IV assumptions is due to the uncontrolled correlation between the cutoff and the outcome variables. This bias is very different from the bias due to the failure to the local randomization design at each window.

In some settings, it is reasonable to assume randomness and the exclusion restriction of cutoffs as potential IVs. Cutoffs assigned to each unit are often arbitrary and irrelevant to factors related to the running variable and the outcome of interest. For example, vaccine policies often assign units to treatment or to the status quo based on a certain cutoff of a continuous running variable, such as age (Bor et al., 2014; Basta and Halloran, 2019; Bermingham et al., 2023; Greene et al., 2022). In such a case, cutoffs vary mostly due to administrative reasons that may not be necessarily related to the distribution of running variables and other potential confounding factors. For example, human papillomavirus (HPV) is recommended for adolescents aged 11–12 years in the United States and those aged 9–13 years or in grades 4–8 in Canada. It is hard to be convinced that having a lower or higher cutoff in one region than the others is related to different age distributions or has a direct effect on the outcome of interest (e.g., sexual behaviors (Smith et al., 2015)). Rather, it is more plausible that different cutoffs in each region might introduce other interventions accompanying the HPV vaccination program that could affect the outcome, which is likely to invalidate the conventional RD inference, but not the proposed IVs. As an another example, Medicaid income eligibility criteria for children differ between states and evolve over time, potentially depending on each state's financial status (De La Mata, 2012); for instance, the eiligibility for children aged 1-18 is set at 261% of the federal poverty level in Rhode Island and 133% for Florida in 2023. As this eligibility cutoff is related to state-level financial status, it is reasonable to assume that different eligibility criteria would not directly affect the health outcomes, even though they could be indirectly associated through state-level infrastructure (e.g., the number of hospitals), which may be accounted for as covariates in the analysis.

In our application study, the cutoffs are not necessarily random because students' transition scores affect their cutoff. Here the cutoffs are the minimum transition scores among students in the town who were eligible for entering the best school. Figure 1(d) suggests that the students from towns with higher cutoffs (e.g., Town 1) tend to have higher transition scores compared to those from towns with lower cutoff. However, as shown in Figure 1(a)-(c) and Table 1, students with a cutoff no larger than 7.56 (i.e., students in Towns 2 and 3) are more likely to have transition scores greater than their cutoff compared to students with a cutoff larger than 7.56 (i.e., students in Town 1). A higher cutoff of Town 1 may be due to a smaller size of the best schools compared to the other two (i.e., being more selective) in addition to overall better performance of students in Town 1. Similarly, students with a cutoff no larger than 7.34 (i.e., students in Town 3) are more likely to be eligible for attending the test school than students with a cutoff larger than 7.34 (i.e., students in Towns 1 and 2). These observations imply that variations in cutoff values satisfy Equation (1).

The cutoff exclusion restriction assumption can also be plausible in our motivating example. A cutoff assigned to each student could be related to the performance of students in the town who could make the best school in the town. It could also be related to the size of the best school relative to other schools in the town as well (the smaller the school size is, the more selective it would be). However, there is no clear reason to believe that these factors directly affect the student's performance on the exam that is taken at the end of high school. In some cases, it might be plausible that a student's performance is easily influenced by their peers' overall performance in the town. Then having higher cutoffs could be associated with the outcome not necessarily through the treatment. This may lead to the violation of the exclusion restriction of cutoffs. We develop a method to test the hypothesis of no effect of going to the best school on baccalaureate scores, while allowing for some degree of violation of the cutoff exclusion restriction.

Unit j	W_{ij}	C_{ij}	D_{ij}	$Z_{ij,1}$	$Z_{ij,2}$	$Z_{ij,3}$	$Z_{ij,4}$	$Z_{ij,5}$
1	0.5	1	0	1	1	1	1	0
2	1.1	1	1	1	1	1	1	1
3	0.7	2	0	1	1	1	0	0
4	2.3	2	1	1	1	1	0	1
5	2.5	3	0	1	1	0	0	0
6	4.1	3	1	1	1	0	0	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$n_i - 3$	1.5	4	0	1	0	0	0	0
$n_i - 2$	6.2	4	1	1	0	0	0	1
$n_i - 1$	3.5	5	0	0	0	0	0	0
n_i	5.1	5	1	0	0	0	0	1

Table 2: A hypothetical example with four nested candidate IVs and one treatment assignment variable in a stratum i.

3.2 Multiple instruments and treatment assignment models

In this section, we establish a formal representation for the random assignment of each proposed IV and a treatment. Then we also illustrate the deviation from this random assignment in the context of sensitivity analysis. The assignment models for proposed IVs and a treatment are particularly useful for representing the nested and conditioning structures among them. They can also be extended to explain the sources of bias that hinder the random assignments of IVs and a treatment.

Let $[m] = \{1, 2, \ldots, m\}$ denote a set of 1 to m. We assume that the observed covariates, denoted by $\mathbf{x}_{ij}$, are controlled by stratification. Consider N units with I different strata.

For each stratum i, there are n_i units with $\sum_{i=1}^{I} n_i = N$. Under the multi-cutoff setting with q distinct cutoffs, potential instruments for unit j from stratum i (hereafter unit ij) for $i \in [I]$ and $j \in [n_i]$ are $Z_{ij,k} = \mathbb{I}(C_{ij} \leq c_{q-k})$ for $k \in [q-1]$. Let $Z_{ij,q}$ denote an indicator of W_{ij} not being less than C_{ij}. Let $\mathbf{Z}_{ij,[q-1]} = (Z_{ij,1}, ..., Z_{ij,(q-1)})$. Table 2 illustrates a hypothetical example with five cutoffs, $\mathcal{C} = \{1, 2, 3, 4, 5\}$. Let $\mathbf{A}_{ij} = \mathbf{Z}_{ij,[q-1]}$ denote a set of $(q-1)$ binary variables. By definition, potential IVs in $\mathbf{A}_{ij}$ are nested and $Z_{ij,q}$ is equivalent to D_{ij} in a sharp RD design. Let $\mathbf{A}^{+}_{ij}$ denote a length-q vector of $\mathbf{Z}_{ij,[q]}$, containing both potential IVs and the actual treatment assignment. Let $\mathbf{A}_{ij,-k}$ denote a length-$(q-2)$ vector of $(Z_{ij,1}, ..., Z_{ij,(k-1)}, Z_{ij,(k+1)}, ..., Z_{ij,(q-1)})$.

With the nested structure, our IV assignment vector $\mathbf{A}_{ij}$ can take q possible values so that we have $2q$ possible assignments in $\mathbf{A}^{+}_{ij}$, because D_{ij} is either 0 or 1. Let $\mathcal{A}^{+}$ denote the set of collection of all these $2q$ vectors. Let $r^{\mathbf{a}}_{ij}$ denote the potential outcome with $\mathbf{a} = (a_1, a_2, \cdots, a_q) \in \mathcal{A}^{+}$ and $\mathbf{r}_{ij} = \{r^{\mathbf{a}}_{ij},\ \mathbf{a} \in \mathcal{A}^{+}\}$ denote a collection of all possible potential outcomes. Our null hypothesis is then $H_0 : r^{d=1}_{ij} = r^{d=0}_{ij}$ (equivalently, $H_0 : r^{a_q=1}_{ij} = r^{a_q=0}_{ij}$ in a sharp RD design) for all units ij. To test this hypothesis with the whole vector in $\mathcal{A}^{+}$, we require a certain assumption that restricts the way that potential IVs in $\mathbf{A}_{ij}$ affect the outcome. Definition 3 introduces the exclusion restriction for a set Q^{+} that contains valid IVs.

Definition 3 *Let $Q^{+} \subseteq [q]$ and $Q = Q^{+} \cap [q-1]$. The reinforced unordered partial exclusion restriction holds for Q^{+} if each unit ij has two potential outcomes depending on the value of D_{ij}, i.e., $r^{d=1}_{ij}$ if $D_{ij} = 1$ and $r^{d=0}_{ij}$ if $D_{ij} = 0$, when conditioning by $(Z_{ij,k})_{k \notin Q}$.*

The above assumption contains “reinforced” feature as Q^{+} may contain the actual treatment assignment in addition to IVs (Karmakar et al., 2021). Additionally, it is “unordered” as the assumption holds by conditioning on the order-free IV sets, $(Z_{ij,k})_{k \notin Q}$. However, unlike the exclusion restriction considered in Zhao et al. (2022), we do not condition on the final component of evidence factors, $Z_{ij,q}$, even if $q \notin Q^{+}$.

Under the reinforced unordered partial exclusion restriction, we propose constructing $|Q^{+}|$ (out of q) number of evidence factors for testing the null of $H_0 : r^{d=1}_{ij} = r^{d=0}_{ij}$ for all units ij. From now we will formally state that each IV k or a direct comparison of treated to control units is *valid* if and only if $k \in Q^{+}$. For example, suppose that there are five cutoffs, i.e., $\mathcal{C} = \{c_1, \ldots, c_5\}$ and $Q^{+} = \{2, 3, 5\}$. This implies that two proposed IVs, i.e., $Z_{ij,2} = \mathbb{I}(C_{ij} \leq c_3)$ and $Z_{ij,3} = \mathbb{I}(C_{ij} \leq c_2)$, and a direct comparison, i.e., $Z_{ij,5} = \mathbb{I}(W_{ij} \geq C_{ij})$, are valid. Each of the second and third proposed IVs is assumed to affect the outcome only through D_{ij}, conditioning on $Z_{ij,1}$ and $Z_{ij,4}$. Similarly, a direct comparison with $Z_{ij,5}$ directly affects the outcome conditioning on $Z_{ij,1}$ and $Z_{ij,4}$. The order of three analyses with $Z_{ij,2}, Z_{ij,3}$, and $Z_{ij,5}$ would not affect the results (i.e., order-free). The three inferential outcomes (i.e., p-values) derived from each analysis could provide nearly orthogonal pieces of evidence for the same null hypothesis.

Next, consider the randomization-based inference method. We start by proposing a general assignment model for the IV and treatment assignments which allows for unmeasured confounding in each assignment. This assignment model is motivated by sensitivity analysis in Rosenbaum (2002). Let $u_{ij,k}$ with $0 \leq u_{ij,k} \leq 1$ denote an unmeasured covariate that hinders a random assignment of $Z_{ij,k}$ for $k \in [q]$, even after stratification on observed covariates $\mathbf{x}_{ij}$. Then with $\mathcal{F} := \{(\mathbf{r}_{ij}, \mathbf{x}_{ij}, u_{ij,k}) : i \in [I],\ j \in [n_i],\ k \in [q]\}$, we consider the following IV assignment of $Z_{ij,k}$ with κ_k being an arbitrary function. For the convenience in notation, let $Z_{ij,0} \equiv 1$. Then for $i \in [I],\ j \in [n_i],\ k \in [q-2]$:

$$\begin{aligned} Pr(Z_{ij,k} = 1 | \mathcal{F}, \mathbf{A}_{ij,-k}) &= Pr(Z_{ij,k} = 1 | \mathcal{F}, Z_{ij,k-1}, Z_{ij,k+1}) \\ &= \mathbb{I}(Z_{ij,k-1} = 1, Z_{ij,k+1} = 0) \frac{\exp\{\kappa_k(\mathbf{x}_{ij}) + \gamma_k u_{ij,k}\}}{1 + \exp\{\kappa_k(\mathbf{x}_{ij}) + \gamma_k u_{ij,k}\}} \\ &\quad + \mathbb{I}(Z_{ij,k-1} = 1, Z_{ij,k+1} = 1); \qquad (2) \\ Pr(Z_{ij,q-1} = 1 | \mathcal{F}, \mathbf{A}_{ij,-(q-1)}) &= \mathbb{I}(Z_{ij,q-2} = 1) \frac{\exp\{\kappa_{q-1}(\mathbf{x}_{ij}) + \gamma_{q-1} u_{ij,q-1}\}}{1 + \exp\{\kappa_{q-1}(\mathbf{x}_{ij}) + \gamma_{q-1} u_{ij,q-1}\}}. \end{aligned}$$

The first line of Equation (2) is due to the nested structure in $\mathbf{A}_{ij}$, i.e., $Z_{ij,k}$ is conditionally independent of $Z_{ij,k'}$ for all $k' \neq k-1, k+1$, and $Z_{ij,k} = 1$ only if $Z_{ij,k-1} = 1$. A parameter γ_k in (2) quantifies the influence of unmeasured covariates $u_{ij,k}$ that is present after conditioned on $\mathbf{x}_{ij}$ when $Z_{ij,k-1} = 1$ and $Z_{ij,k+1} = 0$. Roughly speaking, if $\mathbb{I}(c_{q-k-1} < C_{ij} \leq c_{q-k})$ is not random conditioning on the observed covariates, γ_k would be non-zero.

On the other hand, one of our proposed evidence factors is a direct comparison between the treated and control units. Let $\mathcal{W}(C_{ij})$ denote the window around each standardized cutoff that should include zero, and this window may vary depending on a value of C_{ij}. For example, in cases where the running variable exhibits dispersed values (e.g., students from Town 3 in Figure 1(d)), we may consider a wider window around the cutoff within which a random assignment to the treatment is assumed. Then the treatment assignment with $Z_{ij,q}$ is as follows.

$$Pr(Z_{ij,q} = 1 \mid \mathcal{F}, \mathbf{A}_{ij}, W_{ij} - C_{ij} \in \mathcal{W}(C_{ij})) = \frac{\exp\{\kappa_q(\mathbf{x}_{ij}) + \gamma_q u_{ij,q}\}}{1 + \exp\{\kappa_q(\mathbf{x}_{ij}) + \gamma_q u_{ij,q}\}}. \qquad (3)$$

Here a non-zero γ_q implies the presence of unmeasured covariate that renders treatment assignment non-random within each window even after conditioning on the observed covariates and $\mathbf{A}_{ij}$. Conditioning on $\mathbf{A}_{ij}$ is essentially conditioning on C_{ij}; once we know the values of the whole (q-1) potential IVs, i.e., $\mathbf{A}_{ij}$, C_{ij} is known. Therefore, the assignment model (3) is analogous to the local randomization design using a standardized running variable, assuming a single cutoff: $Z_{ij,q} = 1$ if and only if $D_{ij} = 1$ regardless of the cutoff. When the adjusting covariates $\mathbf{x}_{ij}$ include the running variable then the window $\mathcal{W}$ can be implied by imposing a *caliper* on the running variable (e.g., setting that the difference in W_{ij} among the units within the window is no greater than 0.2 at each cutoff). Note that unmeasured covariates from the IV assignment model in (2) and from the treatment assignment model in (3), i.e., $u_{ij,[q-1]}$ and $u_{ij,q}$, do not necessarily induce bias in other analyses. A non-random assignment of the actual treatment variable would not affect the validity of an IV (i.e., not affecting the reinforced unordered partial exclusion restriction). Moreover, by conditioning on $\mathbf{A}_{ij}$ in the assignment model (3), any bias in potential IVs would not affect the validity of a direct comparison.

4. Reinforced design with evidence from multiple cutoffs

4.1 Constructing evidence factors

Given the assignment models of (2) and (3), we propose building $|Q^+|$ many of evidence factors as follows for testing the sharp null hypothesis $H_0 : r_{ij}^{d=1} = r_{ij}^{d=0}$ when the reinforced unordered partial exclusion restriction holds for Q^+. For now let us assume that $\gamma_k = 0$ for all $k \in Q^+$ so that the assignment of each IV or the treatment is not biased. We will relax this restriction later. For each $k \in Q$, obtain a p-value P_k for testing the sharp null hypothesis by conditioning on $\mathbf{A}_{ij,-k}$ in addition to the observed covariates. If $q \in Q^+$, obtain a p-value P_q by testing the same null hypothesis while conditioning on $\mathbf{A}_{ij}$ given the window where the local randomization is assumed to hold. For each analysis, given the models (2) and (3), we perform randomization-based inference, which leverages randomization mechanisms of IVs or a treatment within strata. We consider testing the RD causal effect using an one-sided non-parametric test, such as stratified Wilcoxon rank-sum statistics and Hodges-Lehman aligned rank statistics. As a result of $|Q^+|$ number of analyses, we obtain $\{P_k : k \in Q^+\}$. The following lemma demonstrates that each analysis is valid (i.e., controlling Type-I error under the null) regardless of the invalidity of other analyses.

Lemma 4 *Suppose that the reinforced unordered partial exclusion restriction holds for Q^+ and $|Q^+| \geq v$ with $v \geq 1$ and $\gamma_k = 0$ for all $k \in Q^+$. Then we have $Pr(P_k \leq p_k) \leq p_k$ for $p_k \in [0,1]$ under the null H_0 regardless of other invalid instruments within the candidate set in $\mathbf{A}_{ij}$ or a direct comparison.*

Then under the regularity conditions on the outcome outlined in Zhao et al. (2022) (see Theorem 4.1), Theorem 5 demonstrates the nearly independence properties of p-values among $\{P_k : k \in Q^+\}$.

Theorem 5 *Suppose that the regularity conditions hold and the reinforced unordered partial exclusion restriction holds for Q^+. Then under the assignments models (2) and (3) with $\gamma_k = 0$ for all $k \in Q^+$, p-values from the proposed reinforced design are stochastically larger than the uniform under the null among valid IVs and a direct comparison. In other words, for any $p_k \in [0,1]$,*

$$Pr(P_k \leq p_k, \ \forall k \in Q^+) \leq \prod_{k \in Q^+} p_k$$

Due to the above properties of the p-values, we can easily combine the p-values from cutoff-based IVs and the p-value from the local randomization framework. This allows us to draw comprehensive causal conclusions while avoiding complicated multiple testing procedures. In particular, we can obtain at least v number of nearly independent p-values for any $1 \leq v \leq q$ and have a single, combined p-value. Let $P_{(k)}$ denote the k^{th} order statistic of $(P_1, \dots, P_q)$ and $U_1, \dots, U_v$ are iid Uniform[0,1] random variable.

Corollary 6 *Suppose that the reinforced unordered exclusion restriction is satisfied for a set $Q^+ \subseteq [q]$ with $|Q^+| \geq v$ $(v \geq 1)$. When f is coordinate-wise non-decreasing, satisfying $Pr\{f(U_1, \dots, U_v) \leq \alpha\} \leq \alpha$ for any $0 \leq \alpha \leq 1$, $f(P_{(q)}, \dots, P_{(q-v+1)})$ is a valid p-value. In other words, $Pr\{f(P_{(q)}, \dots, P_{(q-v+1)}) \leq \alpha\} \leq \alpha$.*

The results of Corollary 6 suggest that we can use the v largest p-values, i.e., $\{P_{(q)}, \ldots, P_{(q-v+1)}\}$, among q to construct a single valid p-value. Different methods for combining p-values determine the specifics of the function f. One of the simplest examples includes Fisher's method, where $f(P_{(q)}, \ldots, P_{(q-v+1)}) = -2\sum_{k=1}^{v} \ln(P_{(q-k+1)})$.

4.2 Sensitivity analysis

In RD designs, sensitivity analyses have largely focused on evaluating sensitivity to particular model specifications under the continuity-based design (Cattaneo et al., 2015; Bloom, 2012). For example, different parametric or non-parametric models can be applied to assess how the causal estimates would change if the presumed model is wrong. The impact of a range of observations around the cutoff included in the analysis (i.e., bandwidth size) and variations in the width of window are also commonly considered in sensitivity analyses (Oldenburg et al., 2016; Cattaneo et al., 2015). In our approach, we do not need to posit any parametric models but rely on non-parametric tests (e.g., stratified Wilcoxon rank tests). Therefore, we instead focus on examining sensitivity to the deviation from the IV assumptions for each proposed IV and the random assignment of the treatment near the cutoffs under the local randomization framework.

One of the advantages of evidence factors design is that it allows us to perform sensitivity analyses by varying the parameter of γ_k in the models (2) and (3) ($k \in [q]$). Note that a non-zero value of γ_k in (2) implies a biased assignment of a cutoff c_{q-k} given the observed covariates ($k \in [q-1]$) while a non-zero value of γ_q in (3) implies a biased treatment assignment given the observed covariates and the cutoffs. Given u_{ijk} and γ_k, one could calculate a corresponding p-value for testing the sharp null using the assignment distribution. Since u_{ijk}'s are unknown, we use the *maximum* p-value over u_{ijk} values in [0,1] denoted by $\overline{P}_{k,\Gamma_k}$ when the sensitivity parameter is at most $\Gamma_k \geq 1$. The next Corollary states that we can consider the maximum p-value as a valid p-value and they also are stochastically larger than the uniform distribution under the null using the same conditions as in Theorem 5. This is because non-zero values of $\Gamma_{k'}$ ($k' \neq k, k' \in [q]$) would not necessarily affect the validity of $\overline{P}_{k,\Gamma_k}$, i.e., $Pr(\overline{P}_{k,\Gamma_k} \leq p_k) \leq p_k$ for any $p_k \in [0,1]$ under the null.

Corollary 7 *Suppose that the regularity conditions hold and the reinforced unordered partial exclusion restriction holds for Q^+. Then under the assignments models* (2)*, the results of Theorem 5 and Corollary 6 hold when we replace P_k by $\overline{P}_{k,\Gamma_k}$ for any $k \in Q^+$.*

Based on the above result, we can conduct sensitivity analysis by varying each of sensitivity parameters Γ_k ($k \in [q]$). Specifically, by varying a parameter Γ_k ($k \in [q-1]$), we can examine the changes in our causal conclusion when each IV assignment is biased. Our sensitivity parameter of Γ_q performs the sensitivity analysis that is typically conducted in the local randomization-based framework, with conditioning on $\mathbf{A}_{ij}$. It is important to note that the value of Γ_k (i.e., any departure from the underlying assumption of evidence factor k) would not affect other factor k' ($k \neq k'$). This allows us to assess sensitivity to biases in different directions that do not overlap.

Remark 8 *Under the local randomization framework, we often transform the outcomes while reflecting some (parametric) relationships between potential outcomes and a running* *variable so that within the window, the running variable would not confound the treatment effect (Sales and Hansen, 2019; Cattaneo et al., 2016a). In this case, one can also examine the impact of model specifications on the results, as well as investigating sensitivity to the unmeasured confounding captured by Γ_q. However, a potential source of bias that affects the former model specifications could also affect the local randomization assumption. Consequently, different from sensitivity analyses within $\{\Gamma_k : k \in [q]\}$, the results of sensitivity analyses to model specifications and local randomization are likely to be correlated, affecting each other.*

5. Fuzzy RD settings

Until now we only consider a sharp RD design where the treatment is received if and only if $W_{ij} \geq C_{ij}$ for all units ij. However, a fuzzy RD design is common in practice where having $W_{ij} \geq C_{ij}$ does not necessarily determine the treatment assignment but alters the probability of receiving the treatment. In our application study, for example, we observe that some eligible students, who could have attended the best school in their town, chose to enroll in other schools. In a fuzzy RD design, $Z_{ij,q} := \mathbb{I}(W_{ij} \geq C_{ij})$ is acting like an IV, rather than a treatment variable, as having $Z_{ij,q} = 1$ increases the probability of receiving the treatment but does not definitively determine the treatment assignment. Lemma 9 below demonstrates sufficient conditions for each of $Z_{ij,k}$ ($k \in [q-1]$) to be a potential IV (Recall Definition 1).

Lemma 9 *Suppose that W_{ij} is independent of C_{ij} and $F_W(c_{q-k+1}) > F_W(c_{q-k})$ with F_W denoting a density function of W_{ij}. Assume that Condition (C1) and one of (C2.1) or (C2.2) hold.*

(C1) $Pr(D_{ij} = 1 \mid W_{ij} \geq C_{ij}, C_{ij} = c_k) \geq Pr(D_{ij} = 1 \mid W_{ij} \geq C_{ij}, C_{ij} = c_{k'})$ *for all cutoffs k and k' such that $c_k < c_{k'}$.*

(C2.1) $Pr(D_{ij} = 1 \mid W_{ij} < C_{ij}, C_{ij} = c_k) = 0$ *for all cutoffs $k = [q]$*

(C2.2) Let $\delta_k := Pr(D_{ij} = 1 \mid W_{ij} \geq C_{ij}, C_{ij} = c_k) - Pr(D_{ij} = 1 \mid W_{ij} < C_{ij}, C_{ij} = c_k)$. *Then* $\delta_{k'}/\delta_k \geq Pr(W_{ij} < C_{ij} \mid C_{ij} = c_k)/Pr(W_{ij} < C_{ij} \mid C_{ij} = c_{k'})$ *if $c_k < c_{k'}$ or $\delta_k = 0$ for all $k = [q]$.*

Then $Pr(D_{ij} = 1 \mid Z_{ij,k} = 1) > Pr(D_{ij} = 1 \mid Z_{ij,k} = 0)$ *with* $Z_{ij,k} = \mathbb{I}(C_{ij} \leq c_{q-k})$ *for $k = [q-1]$.*

Note that in a sharp RD setting, (C1) and (C2.1) are satisfied. Condition (C1) states that the probability of receiving the treatment given that the value of a running variable is above the cutoff is not decreasing as the cutoff is increasing. Condition (C2.1) implies that units with the value of a running variable less than cutoff would not receive the treatment, so non-compliance only occurs in one-direction (i.e., one-sided non-compliance). However, this may not hold in many cases where units with the cutoff below the cutoff are able to receive the treatment. In such a case, (C2.2) can be considered instead, where δ_k indicates the difference in probability of accepting the treatment between eligible and ineligible units. The ratio of this difference in a higher cutoff to a lower cutoff should be no less than the proportion of units with the running variable less than the cutoff between these two cutoffs. The corollary below states that under a fuzzy RD design where we can have up-to q potential IVs, evidence factors analysis can be performed.

Corollary 10 *Under the same conditions as in Theorem 5, the results in Theorem 5 hold under a fuzzy RD design with $Z_{ij,k} = \mathbb{I}(C_{ij} \leq c_{q-k})$ and $Z_{ij,q} = \mathbb{I}(W_{ij} \geq C_{ij})$.*

Here the unmeasured confounder $u_{ij,q}$ in the assignment model (3) indicates non-random allocation of units below and above the cutoff rather than non-random treatment assignment.

6. Cluster-level cutoffs settings

In Equation (2), each of the proposed IVs is assigned to a single unit, which implies that the cutoff is also applied individually. However, in many other aforementioned studies, in addition to our motivating study of school allocations, each cutoff is applied to *clusters* of multiple units. For example, students in the same town were assigned the same cutoff value; units within the same administrative districts are governed by the same vaccination eligibility. These cluster-level cutoff assignments do not necessarily undermine the properties of evidence factors but could affect statistical power as the number of permutations in obtaining p-values is reduced in randomization-based inference.

Consider H_i clusters (e.g., towns) in stratum i. Each cluster ih ($h = 1, \ldots, H_i$) has n_{ih} units with $\sum_{h=1}^{H_i} n_{ih} = n_{i\cdot}$ and $\sum_{i=1}^{I} n_{i\cdot} = N$. Given that cutoffs are assigned to each cluster, denoted by $C_{ih\cdot}$, consider a cluster-level binary potential IV, $Z_{ih\cdot,k}$ ($k = [q-1]$) and a set of (q-1) IVs, $\mathbf{A}_{ih\cdot} = \mathbf{Z}_{ih\cdot,[q-1]}$. Then we assign the same $\mathbf{A}_{ih\cdot}$ to n_{ih} units in cluster ih so that $\mathbf{A}_{ihj} = \mathbf{A}_{ih\cdot}$ for all $j \in [n_{ih}]$. Compared to individual-level treatment allocations, this likely reduces the number of possible permutations (e.g., $n_{i\cdot}!$ to $H_i!$ across i). Similarly define the observed and potential outcomes for unit ihj in stratum i and cluster h, i.e., $r^{\mathbf{a}}_{ihj}$ and r^{d}_{ihj} ($j = [n_{ih}]$, $k = [q]$). Our sharp null hypothesis is then $H_0 : r^{d=1}_{ihj} = r^{d=0}_{ihj}$ for all i, h, j. Note that even in the presence of the within-cluster interference (e.g., where peers' school choices affect one's performance in the exam), randomization-based inference is valid to test the sharp null hypothesis (Rosenbaum, 2007).

We have individual-level observed and unobserved confounders, i.e., $\mathbf{x}_{ihj}$ and $u_{ihj,k}$. Let us define $\mathbf{x}_{ih\cdot}$ as a collection of $\{\mathbf{x}_{ihj} : j = [n_{ih}]\}$ and $\overline{u}_{ih\cdot}$ be a univariate summary of $\{u_{ihj} : j = [n_{ih}]\}$. For example, $\overline{u}_{ih\cdot}$ could be a measure of overall peer pressure in town ih in our school admission study, which could be associated with the distribution of cutoffs and the exam scores in the town. It is known that cluster-level treatment (or IV) assignments are *less* likely susceptible to biases from unmeasured confounders compared to individual-level assignments (Hansen et al., 2014). For example, biases could be larger if the treatment assignment (e.g., cutoff assignment) is affected by the individual-level u_{ihj} and assigned to each individual rather than being affected by a collective factor of $\overline{u}_{ih\cdot}$ and assigned to a group of individuals. Let $\mathcal{F} := \{(\mathbf{r}_{ihj}, \mathbf{x}_{ihj}, u_{ihj,k}) : i \in [I],\ h \in [H_i],\ j \in [n_{ih}],\ k \in [q]\}$. Then we have the following cluster-level IV assignment of $Z_{ih\cdot,k}$ with κ_k being an arbitrary function of the vector of the observed covariates from all units in the same cluster. Define

$Z_{ih\cdot,0} \equiv 1$. For $i \in [I],\ h \in [H_i],\ k \in [q-2]$:

$$\begin{aligned} Pr(Z_{ih\cdot,k} = 1 | \mathcal{F}, \mathbf{A}_{ih\cdot,-k}) &= \mathbb{I}(Z_{ih\cdot,k-1} = 1, Z_{ih\cdot,k+1} = 0) \frac{\exp\{\kappa_k(\mathbf{x}_{ih\cdot}) + \gamma_k \overline{u}_{ih\cdot,k}\}}{1 + \exp\{\kappa_k(\mathbf{x}_{ih\cdot}) + \gamma_k \overline{u}_{ih\cdot,k}\}} \\ &\quad + \mathbb{I}(Z_{ih\cdot,k-1} = 1, Z_{ih\cdot,k+1} = 1); \\ Pr(Z_{ih\cdot,q-1} = 1 | \mathcal{F}, \mathbf{A}_{ih\cdot,-(q-1)}) &= \mathbb{I}(Z_{ih\cdot,q-2} = 1) \frac{\exp\{\kappa_{q-1}(\mathbf{x}_{ih\cdot}) + \gamma_{q-1} \overline{u}_{ih\cdot,q-1}\}}{1 + \exp\{\kappa_{q-1}(\mathbf{x}_{ih\cdot}) + \gamma_{q-1} \overline{u}_{ih\cdot,q-1}\}}. \end{aligned}$$

Here γ_k implies the influence of unmeasured, univariate summary of cluster-level covariate $\overline{u}_{ih\cdot,k}$ on the cutoff assignment, specifically whether cluster ih's cutoff is larger or no larger than $q-k$ ($k = [q-1]$). We can interpret these cluster-level IV assignments as replacement of individual-level covariates and IVs in the assignment model in (2) with their corresponding cluster-level variables. It is important to note that neither $\kappa_k(\mathbf{x}_{ih\cdot})$ nor $\overline{u}_{ih\cdot,k}$ for $k \in [q-1]$ necessarily indicate the overall or average values of those variables within cluster ih. Rather, this assignment may be largely driven by a small subset of units within clusters. To illustrate, in our case study, the cutoff of the best school is likely determined by the transition scores of a relatively small subgroup of students around the maximum size of the best school. The scores of students who easily qualify for the best school or fall significantly below the average would have little impact on determining the cutoff.

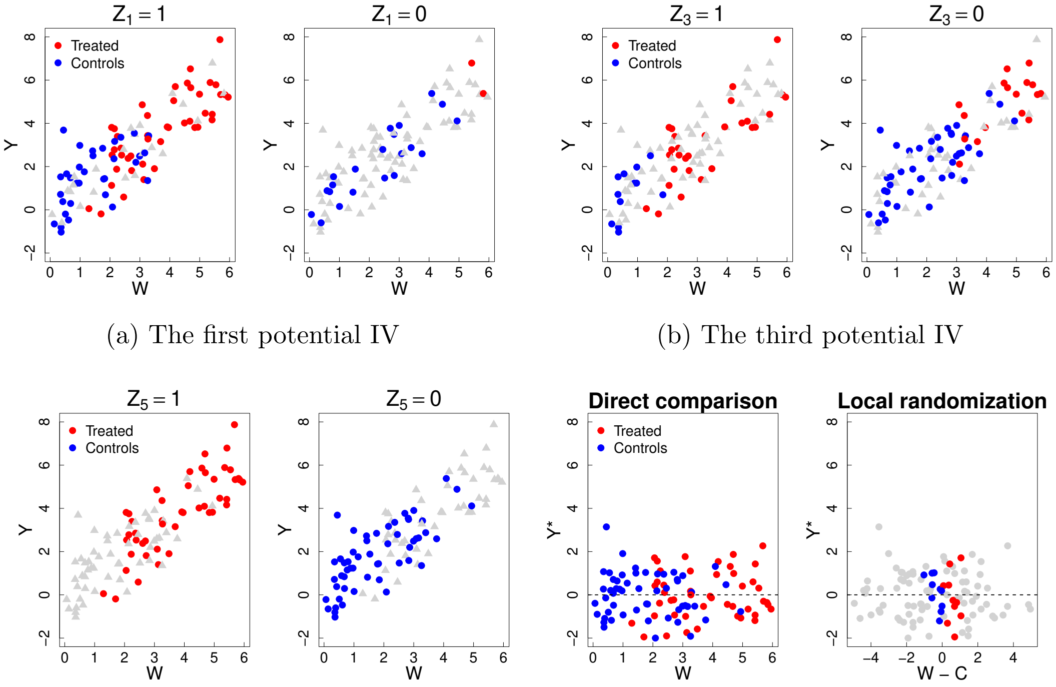
Even though cutoff assignments are cluster-level, the actual treatment assignment is individual-level based on the individual-level running variable, W_{ihj}.

$$Pr(Z_{ihj,q} = 1 \mid \mathcal{F}, \mathbf{A}_{ih\cdot}, W_{ihj} - C_{ih\cdot} \in \mathcal{W}(C_{ih\cdot})) = \frac{\exp\{\kappa_q(\mathbf{x}_{ihj}) + \gamma_q u_{ihj,q}\}}{1 + \exp\{\kappa_q(\mathbf{x}_{ihj}) + \gamma_q u_{ihj,q}\}}.$$

This treatment assignment model is analogous to the assignment model in (3), except that the window is centered around a cluster-level cutoff $C_{ih\cdot}$ rather than an individual-level cutoff. Therefore, each individual ihj is included in the window only if a value of W_{ihj} is close enough to their cluster-specific cutoff, $C_{ih\cdot}$. Once they are in the window, the model above assumes that the probability of receiving the treatment or not depends on individual-level covariates, $\mathbf{x}_{ihj}$ and $u_{ihj,q}$.

7. Simulation studies

In our simulation studies, we examine the performance of the proposed reinforced design under the multi-cutoff RD setting. We generate $N = 1000$ units with five cutoffs, $\mathcal{C} = \{1, 2, 3, 4, 5\}$. We generate $W_{ij} \overset{i.i.d.}{\sim}$ Uniform$(0, 6)$ and $C_{ij} \overset{i.i.d.}{\sim}$ Multinomial$(0.2, 0.2, 0.2, 0.2, 0.2)$. We first consider the sharp RD design where $Pr(D_{ij} = 1 \mid W_{ij}, C_{ij}) = Pr(\mathbb{I}(W_{ij} \geq C_{ij}))$. We also generate an unmeasured variable $U_{ij} := \mathbb{I}(0 \leq W_{ij} - C_{ij} < 0.3)$ to create the violation of the local randomization assumption (also continuity assumption). Let $Z_{ij,k} = \mathbb{I}(C_{ij} \leq c_{5-k})$ for $k = [4]$ and $Z_{ij,5} = \mathbb{I}(W_{ij} \geq C_{ij})$. We do not consider other observed covariates than a running variable for simplicity. Here is the data generating model for the outcome.



(a) The first potential IV

(b) The third potential IV

(c) Actual treatment assignment

(d) Direct comparison with residuals Y^*

Figure 2: (a)-(c): Treatment assignments (red: treated, blue: controls) based on each value of Z_1, Z_3, and Z_5. In each figure titled $Z_k = z$, grey triangle dots denote the units with $Z_k = 1 - z$ $(k = 1, 3, 5;\ z = 0, 1)$ (d): transformed outcomes (residuals) Y^* before and after applying the window constraint.

$$
\begin{aligned}
Y_{ij} &= \sum_{k=1}^{4} \lambda_k Z_{ij,k} + \eta U_{ij} + \beta D_{ij} + W_{ij} + \epsilon_{ij} \qquad (4)\\
&= \sum_{k=1}^{4} \lambda_k \mathbb{I}(C_{ij} \leq c_{q-k}) + \eta \mathbb{I}(0 \leq W_{ij} - C_{ij} < 0.3) + \beta \mathbb{I}(W_{ij} \geq C_{ij}) + W_{ij} + \epsilon_{ij},
\end{aligned}
$$

where $\epsilon_{ij} \sim N(0,1)$. In the above model, a non-zero value for λ_k indicates the violation of the reinforced unordered exclusion restriction of $Z_{ij,k}$ $(k = [4])$. On the other hand, a non-zero value for η implies that having a running variable equal to or larger than the cutoff by 0.3 affects the outcome not through the treatment, D_{ij}. This would invalidate a direct comparison with $Z_{ij,5}$.

Figure 2(a)-(c) illustrate the relationship between the running variable and the outcome, where $\lambda_k = \eta = 0$ $(k = [4])$ from the outcome model (4) using 100 data points. In Figure 2(a), units with rounded dots in the left panel have a value of $Z_{ij,1} = 1$, with the red dots denoting treated units and blue dots denoting the control units. In the right panel, units with $Z_{ij,1} = 0$ have either red and blue dots, depending on their treatment assignment. We observe that the proportion of units receiving the treatment is greater for units with

$Z_{ij,1} = 1$ compared to those with $Z_{ij,1} = 0$. Similar patterns are observed in Figure 2(b). Figure 2(c) demonstrates that all units with $Z_{ij,5} = 1$ are treated, while those with $Z_{ij,5} = 0$ are controls (i.e., a sharp RD).

To test the same null hypothesis of $H_0 : \beta = 0$, we use the stratified Wilcoxon signed rank test for each factor. The analysis with $Z_{ij,k}$ ($k \in [q-1]$) is stratified by $\mathbf{A}_{ij,-k}$ (by exact matching) and W_{ij} (by 1:1 nearest matching). On the other hand, for the the analysis with $Z_{ij,5}$, several strata are generated through the 1:1 nearest matching on the cutoff and the running variable, in addition to the exact matching on $\mathbf{A}_{ij}$. We use the residuals from the linear regression of Y_{ij} on W_{ij} as transformed outcomes for the analysis with $Z_{ij,5}$. This is to reduce the remaining dependency between the outcome and the running variable, ensuring that while the potential outcome within the window may still depend on the treatment assignment, the residuals should not (Sales and Hansen, 2019). Figure 2(d) illustrates the relationship between the running variable W and the residuals Y^* (left panel), and the standardized running variable $W - C$ and the residuals Y^* (right panel). There is no clear linear relationship between W and Y^* compared to the relationship between W and Y in Figures 2(a)-(c).

Under the local randomization, only units within a specified window around the cutoff, which are assumed to be randomly different, are used for RD analyses. The window selection procedure employs the covariate balance test (Cattaneo et al., 2015, 2016a), which identifies the largest window in which all covariates are balanced. In our illustrative and simulation studies, without covariates, we consider a caliper of 0.2 on the running variable value for the last analysis with $Z_{ij,q}$ for illustrative purposes. This ensures that the differences in the running variable value between matched treated and control units do not exceed 0.2, thereby keeping the window width at or below 0.2. The right panel in Figure 2(d) highlights the units that do not meet this caliper criterion in gray. We observe that only units near the zero value of the standardized running variable are considered in the local randomization.

We consider four different cases, (i)-(iv). In case (i), $\boldsymbol{\lambda} = (0, 0, 0, 0)$ and $\eta = 0$, implying that each IV does not have a direct effect on the outcome and there is no effect of having a running variable no less than the cutoff that is not through the treatment. On the other hand, in case (ii), $\eta = 1$, so a direct comparison with Z_{i5} is biased. In case (iii), $\eta = 0$, but $\lambda_4 = 1$, so the fourth IV is biased. Finally, in case (iv), the second and the fourth IVs are invalid and so is a direct comparison. For each case, we combine v largest p-values using Fisher's method and denote the combined p-value given the minimum number of valid factors as $P_{c,v}$.

Table 3 presents the rejection rates using p-values from each factor under four different scenarios. We observe that for each case of violation, the invalidity of one factor does not affect the type-I error of the other factors under the null. For example, in case (iii), except for the results with P_4, the rejection rates of other four factors are all close to or less than $\alpha = 0.05$ level under the null. In general, the power of a direct comparison with P_5 is much greater than that from four IVs (see cases (i) and (iii)). This is expected as the number of units used in the analysis would be much greater with $Z_{ij,5}$ and each of the proposed IV is not as strong as $Z_{ij,5}$. When the combined p-value correctly specify the minimum number of valid factors, its type-I error is no larger than $\alpha = 0.05$. These results support that multiple IVs and a direct comparison provide evidence factors for testing the RD treatment effect.

Rejection rates	P_1	P_2	P_3	P_4	P_5	$P_{c,v}$
(i): $\boldsymbol{\lambda}$ = (0,0,0,0) and η = 0 ($v = 5$)						
β =0.0	0.055	0.052	0.049	0.044	0.014	0.020
0.2	0.098	0.100	0.100	0.061	0.198	0.165
0.4	0.150	0.163	0.164	0.082	0.655	0.647
0.6	0.230	0.250	0.272	0.099	0.952	0.974
0.8	0.334	0.374	0.376	0.119	0.998	1.000
(ii): $\boldsymbol{\lambda}$ = (0,0,0,0) and η =1 ($v = 4$)						
β =0.0	0.050	0.045	0.055	0.043	0.181	0.004
0.2	0.075	0.089	0.088	0.057	0.587	0.029
0.4	0.136	0.149	0.156	0.076	0.912	0.129
0.6	0.199	0.211	0.246	0.094	0.999	0.375
0.8	0.304	0.302	0.331	0.116	1.000	0.669
(iii): $\boldsymbol{\lambda}$ = (0,0,0,1) and η =0 ($v = 4$)						
β =0.0	0.055	0.052	0.052	0.998	0.019	0.022
0.2	0.097	0.100	0.100	0.999	0.219	0.193
0.4	0.149	0.164	0.166	0.999	0.648	0.689
0.6	0.227	0.254	0.268	0.999	0.946	0.977
0.8	0.332	0.371	0.375	0.999	0.997	1.000
(iv): $\boldsymbol{\lambda}$ = (0,1,0,1) and η =1 ($v = 2$)						
β =0.0	0.049	1.000	0.055	0.997	0.223	0.009
0.2	0.075	1.000	0.091	0.997	0.578	0.075
0.4	0.133	1.000	0.156	0.997	0.877	0.172
0.6	0.195	1.000	0.245	0.998	0.983	0.298
0.8	0.306	1.000	0.333	0.998	1.000	0.439

Table 3: Simulation results (rejection rates at $\alpha = 0.05$ based on 1000 independent replicates) for each of the five p-values and the combined P-value, $P_{c,v}$, where v indicates the minimum number of valid analyses for each method.

We also perform the simulation studies under a fuzzy RD design and under cluster-level cutoff settings in the Supplementary Material.

8. Real data application

In this section, we apply our proposed method to evaluate the effect of having access to higher achievement schools on students' performance in a graduation test using the administrative data from Romania (Pop-Eleches and Urquiola, 2013; Bertanha, 2020). We focus on the data from 21 towns with three schools in the cohort of 2001. We focus on the towns with the same number of schools (i.e., three). This is because with heterogeneous number of schools, Equation (1) tends not to hold. For example, suppose that the cutoff of one town with ten schools is lower than that of another town with three schools (even though in fact cutoffs are likely to be higher in towns with more number of schools). Then having a smaller cutoff value would be less likely to lead to a higher chance of attending a better school.

In our context, the running variable W_{ij} denotes the transition score of each student based on which the admission is determined. Our cutoff variable, C_{ij}, denotes the minimum transition score required for admission into the best school of student ij's town, which is in fact cluster-level. The outcome variable, Y_{ij}, is a student's score on the baccalaureate exam. Among the 21 towns, we only consider 13 towns, where the associated cutoff provides a potential IV, satisfying the condition in Definition 1 and Condition (C2.1). There are four towns where schools did not satisfy the condition (C2.1) and those towns are excluded from the analysis as these towns might follow different admission processes. Figure 3 illustrates the distribution of the running variable, the 13 cutoffs, and the outcome variable. It is evident that as students' transition scores increase, they are more likely to enter a better school (denoted by red dots). However, the actual assignment varies depending on students' town's cutoff and their preferences.

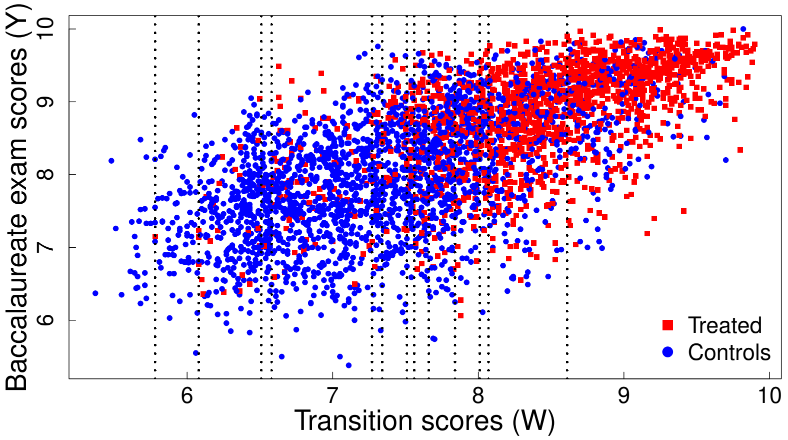


Figure 3: Distribution of the outcome according to the running variable. Each town's cutoff is denoted by 13 vertical lines, and the actual treatment assigned (attending a better school) is denoted by shapes.

Different from the previous RD studies using the same dataset (e.g., Pop-Eleches and Urquiola (2013) and Bertanha (2020)), we consider the effect of attending the best school rather than a better school. In Pop-Eleches and Urquiola (2013) and Bertanha (2020), the treatment is attending a better (not necessarily the best) school and each student's cutoff depends on their transition scores relative to each school's minimum score required to admission. If there are three schools in one town, there could be two "better" schools available to students, and which school is their better school depends on students' transition scores. In our setting, on the other hand, each student's cutoff is set to their town's best school's cutoff. Then each of their outcome could be used in our proposed reinforced design at most twice: one in the IV analyses and the other in a direct comparison. For example, suppose that the cutoffs of the three towns are 8.07, 7.56, and 7.34 (as in the case of Figure 1), and one student ij from Town 2 had her transition score of 7.45. As $\mathbb{I}(C_{ij} \leq c_2) = 1$ and $\mathbb{I}(C_{ij} \leq c_1) = 0$ when $(c_1, c_2, c_3) = (7.34, 7.56, 8.07)$, this student's observation is used in the first IV analysis. If the distance between her score (7.45) and the cutoff (7.56) is considered close enough, then her observation is used in a direct comparison as a control unit who was not admitted to the best school.

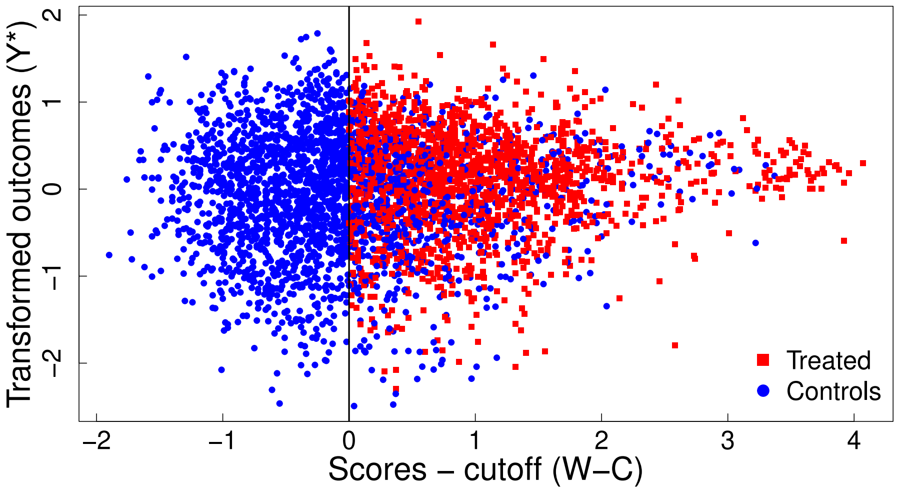


Figure 4: Distribution of the transformed outcome Y^*(residuals from the polynomial model that regressed Y on W) according to the standardized running variable ($W - C$).

In this example, the continuity assumption suggests that the average exam scores would be the same at the cutoff between the students who attended the best school and those who did not if their high school assignments remained the same. On the other hand, the local randomization framework assumes that students whose transition scores are within the window around their town's cutoffs are only randomly different other than their school assignments. Each of these assumptions can be considered reasonable as it is unlikely that students had information about their town's cutoffs before the admission process so they could manipulate their transition scores accordingly. However, the local randomization assumption could be violated especially when the selected window is too wide so that students with the scores far away from the cutoff in either directions could be different on other, unobserved factors (e.g., self-motivation). To reduce the impact of the running variable on the local randomization assumption, we transform the outcome using the residuals from the third-degree polynomial regression model that regresses the outcome on the running variable. Figure 4 presents the distribution of the transformed outcomes on the standardized running variable. We observe that there are control students who did not attend a better school with a higher transition score than the cutoff, but not the other way.

With 13 cutoffs, we can consider twelve potential IVs. The IV assumptions associated with each of twelve IVs suggest that having a specific cutoff would be nearly random and would not have a direct effect on the outcome. These assumptions could be violated when there was a peer effect from other students within a town who were admitted to the best school. However, the town's cutoff is also related to the size of three schools within the town, which is unlikely associated with the outcome. In our analysis, we construct the stratification by performing the 1:1 nearest matching on the running variable. For a particular analysis k, we implement exact matching on $\mathbf{A}_{ij,-k}$ (for $k \in [12]$) or $\mathbf{A}_{ij}$ (for $k = 13$).

i	j	Y_{ij}	D_{ij}	C_{ij}	W_{ij}	$Z_{ij,1}$	$Z_{ij,2}$	$Z_{ij,3}$	$\mathbf{Z}_{ij,[4:12]}$	$Z_{ij,13}$
1	1	9.13	0	7.84	7.94	1	1	1	**0**	1
1	2	8.65	0	8.01	7.94	1	1	0	**0**	0
2	1	8.64	0	7.84	7.16	1	1	1	**0**	0
2	2	8.69	0	8.01	7.51	1	1	0	**0**	0
3	1	9.59	1	7.84	8.80	1	1	1	**0**	1
3	2	8.00	0	8.01	8.80	1	1	0	**0**	1
4	1	8.81	1	7.84	8.17	1	1	1	**0**	1
4	2	9.46	1	8.01	8.18	1	1	0	**0**	1

Table 4: An illustration of the matched pairs (strata) in the analysis of $k = 3$, where $\mathbf{Z}_{ij,[4:12]} = \{Z_{ij,k} : k = 4, 5, \ldots, 12\}$.

By matching on individual-level W_{ij}, for the analysis $k \in [12]$, there is at most one student from each town whose transition scores are similar in each stratum. Table 4 presents four matched pairs (among 181) in the analysis of $k = 3$, where each pair has the same value of $\mathbf{A}_{ij,-3} = (1, 1, 0, 0, \ldots, 0)$ and similar values of W_{ij} within pairs. On the other hand, in the analysis of $k = q$, within each stratum, there are two students from one town, one of whom was eligible for a better school while the other was not. We set the maximum difference in a value of W_{ij} between the matched pair as 0.2, which results in a standardized mean difference (SMD) of W_{ij} between eligible and ineligible students less than 0.8. We investigate the impact of caliper widths on the matching results and the resulting p-value in the Supplementary Material, which shows that a smaller caliper value provides better balance and more insignificant p-value.

$\Gamma_{k,k\in[q]}$	$\overline{P}_{1,\Gamma_1}$	$\overline{P}_{2,\Gamma_2}$	$\overline{P}_{3,\Gamma_3}$	$\overline{P}_{4,\Gamma_4}$	$\overline{P}_{5,\Gamma_5}$	$\overline{P}_{6,\Gamma_6}$	$\overline{P}_{7,\Gamma_7}$	$\overline{P}_{8,\Gamma_8}$	$\overline{P}_{9,\Gamma_9}$	$\overline{P}_{10,\Gamma_{10}}$	$\overline{P}_{11,\Gamma_{11}}$	$\overline{P}_{12,\Gamma_{12}}$	$\overline{P}_{13,\Gamma_{13}}$
1.00	1.000	0.053	1.000	1.000	0.000	1.000	0.000	1.000	0.937	1.000	0.000	0.001	0.148
1.05	1.000	0.114	1.000	1.000	0.000	1.000	0.000	1.000	0.965	1.000	0.000	0.003	0.321
1.10	1.000	0.210	1.000	1.000	0.000	1.000	0.000	1.000	0.981	1.000	0.000	0.006	0.535
1.20	1.000	0.473	1.000	1.000	0.000	1.000	0.000	1.000	0.995	1.000	0.000	0.024	0.869
1.50	1.000	0.967	1.000	1.000	0.025	1.000	0.000	1.000	1.000	1.000	0.006	0.241	1.000
2.00	1.000	1.000	1.000	1.000	0.447	1.000	0.007	1.000	1.000	1.000	0.224	0.825	1.000
2.50	1.000	1.000	1.000	1.000	0.900	1.000	0.150	1.000	1.000	1.000	0.726	0.987	1.000

Table 5: (The maximum) P-values from 13 evidence factors at different values of a sensitivity parameter Γ_k.

Table 5 presents the (maximum) p-values from each of 13 evidence factors at different values of a sensitivity parameter, $\{\overline{\Gamma}_k : k \in [q]\}$. When there is no unmeasured confounder for each factor (i.e., $\Gamma_k = 1.00$), there are four factors resulting in highly significant p-values (< 0.01) and one factor resulting in a moderately significant p-value (0.053). As the value of the sensitivity parameter increases, the maximum p-value for each factor increases. The p-value from a comparison between the students whose scores above and below the cutoff (i.e., $\overline{P}_{13,\Gamma_{13}}$) fails to reject the null at $\Gamma_k = 1.00$. On the other hand, the maximum of p-value from the 7th IV (i.e., having a cutoff no larger than 7.34) remains significant until $\Gamma_k = 2.00$.

Γ_k	$\overline{P}_{c,v=13,\Gamma}$	$\overline{P}_{c,v=12,\Gamma}$	$\overline{P}_{c,v=11,\Gamma}$	$\overline{P}_{c,v=10,\Gamma}$
1.00	0.000	0.000	0.000	0.054
1.05	0.000	0.000	0.002	0.206
1.10	0.000	0.000	0.015	0.468
1.20	0.000	0.001	0.070	0.627
1.50	0.003	0.238	0.683	1.000
2.00	0.631	1.000	1.000	1.000

Table 6: Combined p-values at different values of a sensitivity parameter Γ_k and the number of minimum valid evidence factors, v $(1 \leq v \leq q)$.

Table 6 presents the combined (maximum) p-value from 13 evidence factors, $\overline{P}_{c,v,\Gamma}$, when we assume v minimum number of valid evidence factors given $\Gamma = \{\Gamma_k : k \in [13]\}$. We use the Fisher's method to combine multiple p-values. When there is no unmeasured confounder (i.e., $\Gamma_k = 1.00$ for all $k \in [13]$), we could reject the null hypothesis with combined evidence provided by 13 factors when $v \geq 11$. However, we could not reject the null when we set $v = 10$ at $\alpha = 0.05$. We observe the similar results at $\Gamma_k = 1.05$. At $\Gamma_k = 1.20$, we require at least 12 valid evidence factors out of 13 to reject the null. We observe that as the uncertainties due to unmeasured confounder increase, we require more number of valid factors to obtain significant results. Based on these results, we can conclude that, in the absence of unmeasured confounders, the combined evidence supports a significant effect of attending the best school on the future academic performance. However, as this conclusion is largely driven by a few evidence factors that are highly significant, it is sensitive to the number of valid factors that satisfy the reinforced unordered partial exclusion restriction.

There are several limitations in our data application study. First, there were no individual- nor town-level characteristics available other than students' transition scores. If there were town-level covariates, such as the size of the best school or the number of teachers, that could affect the cutoff assignment, these factors could be used to match different towns instead of matching students. To mitigate the potential bias due to town-level factors, we focused on the town within the same cohort year with the same number of high schools. We match students based on their transition scores instead, assuming that given the transition scores, the assignment to each cutoff (or each town) is random. That is, two units from different towns with similar transition scores (e.g., matched pairs in Table 4) are otherwise exchangeable. This assumption would not hold if there are other town-level characteristics (e.g., the number of teachers available compared to the number of students) that would affect the cutoff and the outcome; our proposed sensitivity analysis quantifies the robustness of the inference to this confounding. Moreover, we only present a set of p-values testing the sharp null hypothesis. The estimates of the additive treatment effect and their corresponding confidence intervals can be obtained using the Hodges-Lehmann estimator based on the Wilcoxon rank sum test. However, as each analysis k is conditioning on different subsets (e.g., those with a cutoff of c_{q-k} or those with a running variable value around the cutoff), the interpretation of the treatment effect estimate from each analysis could vary. It is our potential future research to examine corroborate point estimates across multiple analyses in RD design.

Codes and a sample dataset can be found on github: https://github.com/youjin1207/IVs_inMultiRD.

9. Discussion

This work proposes a new evidence factors design applied to a multi-cutoff RD setting for testing the RD treatment effect. We leverage multiple IVs constructed from multiple cutoffs and combine these potential IVs with a treatment assignment to strengthen our causal conclusions. Our proposed method can also be applied to a fuzzy RD design. While existing literature on RD designs has been relying on the assumptions regarding continuous (or no) changes in certain characteristics at (or near) the cutoffs, our proposed IVs assume the cutoff exclusion restriction of each cutoff. These assumptions can be considered reasonable when different cutoffs are assigned to units nearly arbitrarily or randomly. The evidence factors design also allows us to examine the sensitivity to the cutoff exclusion restriction assumptions. Compared to the literature on multi-level cutoff RD, we utilize the variations in cutoffs rather than standardize them to a constant value and use observations far away from the cutoffs.

In contrast to previous studies that used multiple, different kinds of IVs to construct evidence factors (Karmakar et al., 2021; Zhao et al., 2022), our proposed IVs are all constructed from a cutoff variable. This could easily make bias from each IV analysis concur (e.g., $\boldsymbol{\lambda}$ is likely to be either a vector of all 1's or 0's in our simulation models (4)). However, units with different cutoffs may be subject to different types of bias, particularly when cutoffs are determined by administrative districts, which is common in vaccination eligibility (Bor et al., 2014; Bermingham et al., 2023), political elections (Cattaneo et al., 2016b), and education opportunities (Pop-Eleches and Urquiola, 2013). Then having a smaller or larger value of a cutoff to take the treatment in one district could be biased due to the factors that do not affect the other districts.

As suggested by Donald Campbell in Trochim (1984), our future work includes constructing useful evidence by performing both continuity-based and local randomization approaches that leverages the difference at the cutoff and around the cutoff. In such a case, it would be challenging to disentangle dependency between the statistics from both methods. We can also consider extending our approach to the RD design with multiple running variables that determine the treatment assignments (Papay et al., 2011; Reardon and Robinson, 2012; Díaz and Zubizarreta, 2023).

References

Atila Abdulkadiroğlu, Joshua Angrist, and Parag Pathak. The elite illusion: Achievement effects at boston and new york exam schools. *Econometrica*, 82(1):137–196, 2014.

Robert Ainsworth, Rajeev Dehejia, Cristian Pop-Eleches, and Miguel Urquiola. Why do households leave school value added on the table? the roles of information and preferences. *American Economic Review*, 113(4):1049–1082, 2023.

Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):

444–455, 1996.

Nicole E Basta and M Elizabeth Halloran. Evaluating the effectiveness of vaccines using a regression discontinuity design. *American journal of epidemiology*, 188(6):987–990, 2019.

Charlotte Bermingham, Jasper Morgan, Daniel Ayoubkhani, Myer Glickman, Nazrul Islam, Aziz Sheikh, Jonathan Sterne, A Sarah Walker, and Vahé Nafilyan. Estimating the effectiveness of first dose of covid-19 vaccine against mortality in england: a quasi-experimental study. *American Journal of Epidemiology*, 192(2):267–275, 2023.

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

Howard S Bloom. Modern regression discontinuity analysis. *Journal of Research on Educational Effectiveness*, 5(1):43–82, 2012.

Jacob Bor, Ellen Moscoe, Portia Mutevedzi, Marie-Louise Newell, and Till Bärnighausen. Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiology (Cambridge, Mass.)*, 25(5):729, 2014.

Matias D Cattaneo and Rocio Titiunik. Regression discontinuity designs. *Annual Review of Economics*, 14:821–851, 2022.

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

Matias D Cattaneo, Rocio Titiunik, and Gonzalo Vazquez-Bare. Inference in regression discontinuity designs under local randomization. *The Stata Journal*, 16(2):331–367, 2016a.

Matias D Cattaneo, Rocío Titiunik, Gonzalo Vazquez-Bare, and Luke Keele. Interpreting regression discontinuity designs with multiple cutoffs. *The Journal of Politics*, 78(4): 1229–1248, 2016b.

Raj Chetty, Nathaniel Hendren, and Lawrence F Katz. The effects of exposure to better neighborhoods on children: New evidence from the moving to opportunity experiment. *American Economic Review*, 106(4):855–902, 2016.

Damon Clark. Selective schools and academic achievement. *The BE Journal of Economic Analysis & Policy*, 10(1), 2010.

Julie Berry Cullen, Brian A Jacob, and Steven D Levitt. The impact of school choice on student outcomes: an analysis of the chicago public schools. *Journal of Public Economics*, 89(5-6):729–760, 2005.

Stacy Berg Dale and Alan B Krueger. Estimating the payoff to attending a more selective college: An application of selection on observables and unobservables. *The Quarterly Journal of Economics*, 117(4):1491–1527, 2002.

Dolores De La Mata. The effect of medicaid eligibility on coverage, utilization, and children's health. *Health economics*, 21(9):1061–1079, 2012.

Juan D Díaz and José R Zubizarreta. Complex discontinuity designs using covariates: Impact of school grade retention on later life outcomes in chile. *The Annals of Applied Statistics*, 17(1):67–88, 2023.

Will Dobbie and Roland G Fryer Jr. The impact of attending a school with high-achieving peers: Evidence from the new york city exam schools. *American Economic Journal: Applied Economics*, 6(3):58–75, 2014.

Esther Duflo, Pascaline Dupas, and Michael Kremer. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American economic review*, 101(5):1739–1774, 2011.

Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications*. Chapman and Hall, London), 1996.

David Figlio, Kristian L Holden, and Umut Ozek. Do students benefit from longer school days? regression discontinuity evidence from florida's additional hour of literacy instruction. *Economics of Education Review*, 67:171–183, 2018.

Elisabeth R Gerber and Daniel J Hopkins. When mayors matter: estimating the impact of mayoral partisanship on city policy. *American Journal of Political Science*, 55(2): 326–339, 2011.

Sharon K Greene, Alison Levin-Rector, Emily McGibbon, Jennifer Baumgartner, Katelynn Devinney, Alexandra Ternier, Jessica Sell, Rebecca Kahn, and Nishant Kishore. Reduced covid-19 hospitalizations among new york city residents following age-based sars-cov-2 vaccine eligibility: Evidence from a regression discontinuity design. *Vaccine: X*, 10: 100134, 2022.

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

Ben B Hansen, Paul R Rosenbaum, and Dylan S Small. Clustered treatment assignments and sensitivity to unmeasured biases in observational studies. *Journal of the American Statistical Association*, 109(505):133–144, 2014.

Bikram Karmakar, Dylan S Small, and Paul R Rosenbaum. Reinforced designs: Multiple instruments plus control groups as evidence factors in an observational study of the effectiveness of catholic schools. *Journal of the American Statistical Association*, 116 (533):82–92, 2021.

Jean-William Laliberté. Long-term contextual effects in education: Schools and neighborhoods. *American Economic Journal: Economic Policy*, 13(2):336–377, 2021.

David S Lee. Randomized experiments from non-random selection in us house elections. *Journal of Econometrics*, 142(2):675–697, 2008.

David S Lee and Thomas Lemieux. Regression discontinuity designs in social sciences. *The SAGE handbook of regression analysis and causal inference*, pages 301–27, 2014.

Adrienne M Lucas and Isaac M Mbiti. Effects of school quality on student achievement: Discontinuity evidence from kenya. *American Economic Journal: Applied Economics*, 6 (3):234–263, 2014.

Tatiana Melguizo, Fabio Sanchez, and Tatiana Velasco. Credit for low-income students and access to and academic performance in higher education in colombia: A regression discontinuity approach. *World development*, 80:61–77, 2016.

Brian G Moss and William H Yeaton. Shaping policies related to developmental education: An evaluation using the regression-discontinuity design. *Educational Evaluation and Policy Analysis*, 28(3):215–229, 2006.

Catherine E Oldenburg, Ellen Moscoe, and Till Bärnighausen. Regression discontinuity for causal effect estimation in epidemiology. *Current Epidemiology Reports*, 3:233–241, 2016.

Yasi˙ n Kürşat Önder and Mrittika Shamsuddin. Heterogeneous treatment under regression discontinuity design: Application to female high school enrolment. *Oxford Bulletin of Economics and Statistics*, 81(4):744–767, 2019.

John P Papay, John B Willett, and Richard J Murnane. Extending the regression-discontinuity approach to multiple assignment variables. *Journal of Econometrics*, 161 (2):203–207, 2011.

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

Sean F Reardon and Joseph P Robinson. Regression discontinuity designs with multiple rating-score variables. *Journal of research on Educational Effectiveness*, 5(1):83–104, 2012.

Joseph P Robinson. Evaluating criteria for english learner reclassification: A causal-effects approach using a binding-score regression discontinuity design with instrumental variables. *Educational Evaluation and Policy Analysis*, 33(3):267–292, 2011.

Paul R Rosenbaum. *Observational studies*. Springer, 2002.

Paul R Rosenbaum. Interference between units in randomized experiments. *Journal of the american statistical association*, 102(477):191–200, 2007.

Paul R Rosenbaum. Evidence factors in observational studies. *Biometrika*, 97(2):333–345, 2010.

Paul R Rosenbaum. Some approximate evidence factors in observational studies. *Journal of the American Statistical Association*, 106(493):285–295, 2011.

Paul R Rosenbaum. The general structure of evidence factors in observational studies. *Statistical Science*, 32(4):514–530, 2017.

Adam C. Sales and Ben B. Hansen. Limitless regression discontinuity. *Journal of Educational and Behavioral Statistics*, 45(2):143–174, November 2019. ISSN 1935-1054. doi: 10.3102/1076998619884904. URL http://dx.doi.org/10.3102/1076998619884904.

Leah M Smith, Jay S Kaufman, Erin C Strumpf, and Linda E Lévesque. Effect of human papillomavirus (hpv) vaccination on clinical indicators of sexual behaviour among adolescent girls: the ontario grade 8 hpv vaccine cohort study. *Cmaj*, 187(2):E74–E81, 2015.

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

William MK Trochim. *Research design for program evaluation: The regression-discontinuity approach*, volume 6. SAGE Publications, Incorporated, 1984.

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

Atheendar S Venkataramani, Jacob Bor, and Anupam B Jena. Regression discontinuity designs in healthcare research. *bmj*, 352, 2016.

Anqi Zhao, Youjin Lee, Dylan S Small, and Bikram Karmakar. Evidence factors from multiple, possibly invalid, instrumental variables. *The Annals of Statistics*, 50(3):1266–1296, 2022.