MARGHERITA FORT

University of Bologna, FBK-IRVAPP, CEPR, CESifo and IZA

ANDREA ICHINO

European University Institute, University of Bologna, CEPR, CESifo and IZA

ENRICO RETTORE

University of Padova, FBK-IRVAPP, and IZA

GIULIO ZANELLA

University of Bologna and IZA

MULTI-CUTOFF RD DESIGNS WITH OBSERVATIONS LOCATED AT EACH CUTOFF: PROBLEMS AND SOLUTIONS

January 2022

Marco Fanno Working Papers – 278

Department of Economics and Management "Marco Fanno"
University of Padova
Via del Santo 33, 35123 Padova

Multi-cutoff RD designs with observations located at each cutoff: problems and solutions[*]

Margherita Fort[†] Andrea Ichino[‡] Enrico Rettore[¶] Giulio Zanella[§]

January 14, 2022

Abstract

In RD designs with multiple cutoffs, the identification of an average causal effect across cutoffs may be problematic if a marginally exposed subject is located exactly at each cutoff. This occurs whenever a fixed number of treatment slots is allocated starting from the subject with the highest (or lowest) value of the score, until exhaustion. Exploiting the "within" variability at each cutoff is the safest and likely efficient option. Alternative strategies exist, but they do not always guarantee identification of a meaningful causal effect and are less precise. To illustrate our findings, we revisit the study of Pop-Eleches and Urquiola (2013).

JEL-Code: C01

Keywords: Regression Discontinuity; multiple cutoffs; Normalizing-and-Pooling.

[*] We are grateful to seminar participants at the University of Bologna, University of Padova, the European University Institute, the "Advances in Econometrics - Regression Discontinuity Designs: Theory and Applications" conference at the University of Michigan, the 2018 Ski & Labor Seminar in Landeck, the European Causal Inference Meeting (EuroCIM) in Florence. We also thank Josh Angrist, Erich Battistin, Adriano De Falco, Alice Dominici, Firmin Doko Tchatoka, Greg Kaplan, Chiara Monfardini, Sergio Pastorello, Yannick Reichlin and Jim Snyder for valuable suggestions. All errors are ours.

[†] University of Bologna, FBK-IRVAPP, CEPR, CESifo and IZA; margherita.fort@unibo.it.

[‡] European University Institute, University of Bologna, CEPR, CESifo and IZA; andrea.ichino@eui.eu.

[¶] University of Padova, FBK-IRVAPP, and IZA; enrico.rettore@unipd.it.

[‡] University of Bologna and IZA; giulio.zanella@unibo.it.

1 Introduction

Regression Discontinuity designs with multiple cutoffs are increasingly frequent in the social sciences. Consider a treatment that is provided at different points in space or time (or both), such as an educational program delivered every year at different schools. We will refer to these points of treatment delivery as the "sites". In each site $j \in \{1, \ldots, J\}$, exposure to treatment for individual $i \in \{1, \cdots, N_j\}$ is determined only by the value of a predetermined individual score, X^*_{ij}, relative to a site-specific cutoff, c_j, that may vary across sites. If $X^*_{ij} \geq c_j$ then individual i receives the treatment in site j. In such multi-cutoff setting, it is common to normalize the score variable by taking the difference between the individual score and the site-specific cutoff, $X_{ij} = X^*_{ij} - c_j$. It is then possible to pool all the observations around the unique zero-normalized cutoff and to estimate an overall RD treatment effect using the score-distance from this cutoff, X_{ij}, as the assignment variable. In their illustration of this procedure, Cattaneo et al. (2016) label it as the "Normalizing-and-Pooling" (NP) estimator and document its diffusion in the RD literature. Pooling across sites is appealing when the goal is to estimate an average treatment effect across sites, and becomes the only feasible strategy when the number of subjects per site is too small for a site-specific analysis. This paper studies a setting in which NP may be problematic and in which other pooling strategies may be preferable.

Specifically, we analyze multi-cutoff RD designs in which each cutoff is the value of the running variable for a marginal subject exposed to treatment, so that there is one observation located exactly at each threshold. This situation typically arises when the number of available treatment slots is set ex-ante in each site and, ex-post, this number turns out to be smaller than the number of applicants to each site. Rationing is resolved by first ranking applicants according to a score and then offering treatment down the ranking until all slots are filled. Typically, as a consequence of this allocation rule, the score of the marginally treated subject in a site is chosen as the site's cutoff, and a researcher would observe a disturbing violation of the continuity of the pooled density of the running variable at the unique zero-normalized cutoff. Updating the review in the appendix to Cattaneo et al. (2016), Table A–1 reviews a number of articles (most of which are published in leading journals in economics and political science) that feature an RD design with multiple cutoffs *and* the presence of one subject at each cutoff. Far from being exhaustive, this list indicates that the setting considered in this paper is not a rarity in empirical applications and that the problem and solutions that we study are quite relevant for applied researchers.

Our contribution is to first show that in this setting the estimand of the standard NP estimator *may not* coincide with any meaningful causal parameter, even if the identification assumptions of the sharp RD design (Hahn et al., 2001) hold at each single cutoff. We then provide a framework to study and compare, in terms of bias and precision, the most promising alternatives that have been adopted in the applied literature. Our analysis shows that when the allocation rule results in an observation located exactly at each cutoff, a Site Fixed Effect (SFE) estimation strategy ensures the identification of a meaningful causal parameter. In addition, a SFE estimator may also offer an efficiency advantage. The reason for this advantage is interesting given that fixed effect strategies are typically not so efficient. In a standard regression setting with clustered observations, conditioning on a set of fixed effects for each cluster has contrasting consequences on standard errors: on the one hand, it reduces the variance of the error term; on the other hand, only the within-cluster variability can be used to estimate the treatment effect. Whether the net outcome is a gain or a loss in precision is an empirical matter. In the RD setting that we study, after deriving a sufficient condition for SFE to be *more precise* than NP, we show that this condition is likely to be met. The reason is that a multi-cutoff RD setting is similar to a stratified RCT with a constant probability of treatment in each stratum; therefore, the remarks of Athey and Imbens (2017) apply, i.e., the fixed effect transformation does not reduce in a relevant way the variability of the treatment dummy, while reducing the noise.

We also consider alternatives to a SFE strategy whose estimand, under certain conditions, is close to a meaningful causal parameter. However, these options may themselves be problematic. A framework to compare these alternative strategies with the SFE estimator is missing in the literature and we fill this gap. For example, we show that the strategy of dropping the observations located exactly at the cutoffs and then applying the NP estimator, leads to a potentially large "bias"[1] and is less precise than SFE. Another class of

[1] Here and in the rest of the paper, with a slight abuse of language we use the word "bias" to refer to the

possible strategies is based on re-defining the cutoff without dropping marginal subjects, which may be attractive when the number of observations per site is small. Within this class, an appealing solution adopted by, e.g., Boas and Hidalgo (2011) and Kendall and Rekkas (2012) and which we label as "symmetric-distance" (SYM), consists of setting the cutoff that is relevant for the treated units (resp. non-treated) as the score of the marginally non-treated unit (resp. treated). It is then possible to apply the standard NP estimator using as a running variable the distance from the respective re-defined cutoff. Along similar lines, an even simpler solution (which we label as "SPLIT") is to set the cutoff at half the distance between the scores of the last treated and the first non-treated subjects, and then proceed with the NP strategy. Our empirical analysis shows that both SYM and SPLIT yield estimates that may be close to those generated by the SFE estimator, but they are much less precise. Finally, solutions based on re-defining the running variable in terms of ranks (RK) perform poorly in terms of both "bias" and precision, unless the sample size per site is extremely large.

Our analysis is illustrated empirically by revisiting Pop-Eleches and Urquiola (2013), who use RD to estimate the effects of being admitted to selective high schools in Romania. Thanks to the very large number of students applying to each school, these authors' data set allows us to evaluate at different sample sizes the estimators that we study. Specifically, we implement the following thought experiment in the multi-cutoff RD setting faced by these authors: what would a researcher have obtained by applying alternative estimation strategies to a sequence of hypothetical data sets featuring the same number of schools but a progressively smaller number of randomly chosen applicants per school? In the universe of applicants, Pop-Eleches and Urquiola (2013) apply a parametric SFE strategy after dropping observations located at the cutoffs. This estimator performs well for the causal parameter they are interested in, i.e., the Average Treatment effect on the Treated (ATT), and would have performed equally well had they kept the marginal subjects in the sample. However, we show that a researcher who had omitted the site fixed effects would have incurred a substantial "bias" at smaller sample sizes, with or without marginal subjects in the analysis.

As the range of applications of the RD design expands, an increasing number of prac-

difference between an estimand and a population parameter of interest.

titioners are likely to face settings where the problem that we study is relevant. These practitioners may find of some practical use our discussion of the anatomy of the problem, as well as our analysis of the available strategies that may be adopted to address it.

The rest of the paper proceeds as follows. Section 2 briefly reviews related studies. Section 3 describes the non-problematic multi-cutoff design studied by Cattaneo et al. (2016), which facilitates the description of the problematic multi-cutoff case in Section 4. The safest option that we recommend to handle the problem is based on a Site Fixed Effect estimation strategy that is discussed in Section 5. Section 6 and Section 7 discuss the pros and cons of alternative solutions. Section 8 illustrates our analysis empirically using the data of Pop-Eleches and Urquiola (2013). Section 9 concludes by summarizing lessons for practitioners.

2 Closely related literature

The problem analyzed in this paper is reminiscent of the marginal subject problem studied by de Chaisemartin and Behaghel (2020), which is generated by the same allocation rule that we consider. These authors show that observations located at each cutoff should be dropped, for a very specific reason that is orthogonal and complementary to our setting. Namely, subjects are characterized by an intrinsic type: being an “accepter” or a “refuser” of a potential treatment offer. While the proportion of accepters is continuous around a given threshold and may differ from 1 in a fuzzy RD, it must be exactly 1 for subjects located at the threshold because these subjects, by construction, must have accepted an offer.

The Simpson’s paradox may also appear to be related to our analysis. This paradox is a statistical phenomenon whereby the relationship between a cause and an effect holds with a certain sign in a population P and, at the same time, holds with the opposite sign in J non-overlapping groups P_j, such that $P_1 \cup P_2 \cup \cdots \cup P_J = P$, that compose the population (Pearl, 2014). However, the problem that we study, being essentially a problem of weighting the different groups appropriately, would manifest itself even in a context in which the sign reversal that defines the paradox does not occur.

Similarly, the problem discussed by Barreca et al. (2015), where the “donut” solution works, is intrinsically different from the one that we study. In their setting, observations at exactly zero distance from the cutoff are problematic only because of heaping and manipulation of the running variable, which we rule out.

Our contribution is most closely related to Cattaneo et al. (2016) and Bertanha (2020), two articles that study the implications of multiple cutoffs for the interpretation of RD estimates. Bertanha (2020) focuses on how one can exploit the multiple cutoffs to generalize local average treatment effects identified and estimated at observed cutoffs, and to average treatment effects over general counterfactual distributions of individuals. Cattaneo et al. (2016) focus instead on identification, estimation, and formal interpretation of a pooled treatment effect via the NP approach. The setting of Cattaneo et al. (2016) is different from ours because they study RD designs in which each threshold is exogenously determined and so the probability that a subject is located exactly at the cutoff is zero. Such non-problematic setting is illustrated next.

3 The non-problematic multi-cutoff design

A program is delivered in $J \geq 2$ different sites, indexed by $j \in \{1, \dots, J\}$, where J is the number of sites in the population. None of the results that follow depends on J, which is kept fixed throughout the analysis. Site j receives applications from N_j individuals indexed by i. Each applicant is attached to one (and only one) site and our analysis is conditional on the selection process that matches applicants and sites. Applicants receive the treatment if and only if they are allocated a slot. In each site j, slots are allocated according to the following rule:

Allocation rule 1 *In each site j:*

1.a The N_j applicants are ranked according to a continuous pre-treatment score X^*_{ij}, with density $f_{X^*}(x^*|j)$ and c.d.f. $F_{X^*}(x^*|j)$.

1.b Those applicants whose score X^*_{ij} is greater than a predetermined cutoff $c_j \in \{c_1, c_2, ..., c_J\}$ are offered admission to the program.

As a result of this allocation rule, the treatment indicator is $D_{ij} = \mathbb{I}[X^*_{ij} \geq c_j]$. Let $Y_{ij}(D_{ij}) \in \{Y_{ij}(0), Y_{ij}(1)\}$ denote potential outcomes for individual i in site j when, respectively, she is not exposed or she is exposed to the treatment. Thus, the individual treatment effect is given by

$$\tau_{ij} \equiv Y_{ij}(1) - Y_{ij}(0), \tag{1}$$

while the outcome observed by the researcher is

$$Y_{ij} = Y_{ij}(1)D_{ij} + Y_{ij}(0)(1 - D_{ij}). \tag{2}$$

Like Cattaneo et al. (2016), we maintain the following three standard assumptions that guarantee the internal validity of a sharp RD design in each site.[2] The first ensures the existence of a full discontinuity in the exposure to treatment at each cutoff (first stage):[3]

Assumption 1

$$\lim_{\varepsilon \to 0^+} \mathbb{E}[D_{ij}|X^*_{ij} = c_j + \varepsilon, j] = 1 \tag{3}$$

$$\lim_{\varepsilon \to 0^+} \mathbb{E}[D_{ij}|X^*_{ij} = c_j - \varepsilon, j] = 0. \tag{4}$$

The second guarantees the continuity of potential outcomes around the respective cutoffs:

Assumption 2 *$\forall j$, $\mathbb{E}[Y_{ij}(0)|X^*_{ij}, j]$ and $\mathbb{E}[Y_{ij}(1)|X^*_{ij}, j]$ are continuous in X^*_{ij} at cutoff c_j.*

The third guarantees that the density of the running variable in each site is also continuous around the site-specific cutoff:

Assumption 3 *The density $f_{X^*}(x^*|j)$ of the running variable X^*_{ij} is positive and continuous at any x^*, specifically at each cutoff c_j.*

In such a setting, researchers may be interested in estimating an average treatment effect across all sites. If N_j is sufficiently large in every site, an obvious way to proceed is the following: (i) employ a separate RD design in each site to estimate the site-specific estimand under Allocation rule 1,[4]

$$\tau_j^{RD_1} \;=\; \lim_{\varepsilon \to 0^+} \mathbb{E}[Y_{ij}|X^*_{ij} = c_j + \varepsilon, j] - \lim_{\varepsilon \to 0^+} \mathbb{E}[Y_{ij}|X^*_{ij} = c_j - \varepsilon, j], \tag{5}$$

[2] To simplify the exposition, we focus here on a Sharp RDD. We discuss the Fuzzy case in footnote 9.

[3] The conditioning on j on the right-hand side of equations (3) and (4) is shorthand notation for "individual i is attached to site j", and is used extensively in what follows.

[4] Here and in what follows we will focus only on the population estimands of the different estimation strategies. The estimators can be easily derived, as usual, by replacing the population moments with their sample analogs.

which, given Assumptions 1 and 2, coincides with the site-specific Average Treatment Effect at the cutoff, $\tau_j^{ATE} \equiv \mathbb{E}[\tau_{ij}|j, X_{ij}^* = c_j]$; (ii) take an average across sites of the site-specific $\tau_j^{RD_1}$'s with the preferred weights. However, in applied research it is often the case that, even if Assumptions 1, 2 and 3 are satisfied, the effective number N_j of observations available to the researcher in some sites may not be large enough to allow for a reliable within-site RD estimation. In what follows, we restrict the analysis to situations of this kind, in which researchers have no choice and *can only* estimate an average treatment effect across sites.

In this case, a simple solution is the "Normalizing and Pooling" estimator described by Cattaneo et al. (2016). After defining the normalized running variable $X_{ij} = X_{ij}^* - c_j$, with distribution $f_X(x_{ij}|j)$, treatement exposure is determined by a unique cutoff that is equal to zero in all sites: $D_{ij} = \mathbb{I}[X_{ij} \geq 0]$. Pooling all observations around this unique cutoff, the estimand of a single RD design for all sites is the NP estimand under Allocation rule 1:

$$\tau^{NP_1} \;=\; \lim_{\varepsilon \to 0^+} \mathbb{E}[Y_{ij}|X_{ij} = \varepsilon] - \lim_{\varepsilon \to 0^+} \mathbb{E}[Y_{ij}|X_{ij} = -\varepsilon]. \tag{6}$$

To simplify the notation, we denote open neighborhoods below and above the zero cutoff as

$$\begin{aligned} \mathbb{O}^- &= \{X_{ij} : -\varepsilon < X_{ij} < 0\} \quad \text{with} \quad \varepsilon \to 0^+; \\ \mathbb{O}^+ &= \{X_{ij} : 0 < X_{ij} < +\varepsilon\} \quad \text{with} \quad \varepsilon \to 0^+, \end{aligned} \tag{7}$$

so that τ^{NP_1} can be written as

$$\begin{aligned} \tau^{NP_1} &= \mathbb{E}[Y_{ij}|\mathbb{O}^+] - \mathbb{E}[Y_{ij}|\mathbb{O}^-] \\ &= \sum_{j=1}^{J} \left[\mathbb{E}[Y_{ij}|\mathbb{O}^+, j]\; P[j|\mathbb{O}^+] - \mathbb{E}[Y_{ij}|\mathbb{O}^-, j]\; P[j|\mathbb{O}^-] \right], \end{aligned} \tag{8}$$

where $P[j|\mathbb{O}^+]$ is shorthand notation for the probability $P[j|X_{ij} \in \mathbb{O}^+]$ that a randomly drawn observation whose normalized running variable takes a value in $\mathbb{O}^+$ comes from site j, and similarly for $P[j|\mathbb{O}^-]$. Note that the summation on j runs over all the J sites in the population. We will adopt this convention in all the analogous expressions that follow.

Under Assumption 3, it is

$$P[j|\mathbb{O}^+] = P[j|\mathbb{O}^-] = P[j|0] = \frac{1}{J}\frac{N_j f_j}{(Nf)}, \tag{9}$$

where $f_j = f_X(0|j)$ is the density of the running variable at the cutoff in site j and $\overline{(Nf)}$ is the average across sites of $N_j f_j$. The thought experiment that we use here to characterize the weights, and that will be used repeatedly in what follows, is that in a hypothetical repeated sampling process the number of applicants in each site N_j is kept fixed. It is then easy to see that the estimand τ^{NP_1} coincides with the Average causal Effect of Treatment at the cutoff across all sites and observations under Allocation rule 1 (ATE_1), i.e.,

$$\tau^{ATE_1} \quad = \quad \mathbb{E}\left[\mathbb{E}[\tau_{ij}|0,j]\right] = \sum_{j=1}^{J} \left[\mathbb{E}[Y_{ij}(1)|0,j] \; - \mathbb{E}[Y_{ij}(0)|0,j]\right] \; P[j|0]. \tag{10}$$

Therefore, in the non-problematic setting of Allocation rule 1, the Normalizing-and-Pooling approach allows the researcher to estimate the corresponding ATE_1. It is also important to note that, different from the problematic setting of Allocation rule 2 that we discuss next, in this setting the ATE_1, the ATT_1 and the $ATNT_1$ are all equal to the single causal effect τ^{ATE_1}, because of the equality of weights in equation (9).[5] Therefore,

$$\tau^{ATE_1} = \tau^{ATT_1} = \tau^{ATNT_1} = \tau^{NP_1}. \tag{11}$$

4 The problematic multi-cutoff design

Maintaining Assumptions 1, 2 and 3, consider the following alternative allocation rule, that differs from Allocation rule 1 because of Part (2.b):

Allocation rule 2 *In each site j:*

(2.a) The N_j applicants are ranked according to a continuous pre-treatment score X^*_{ij}, with density $f_{X^*}(x^*|j)$ and c.d.f. $F_{X^*}(x^*|j)$.

(2.b) The number K_j of available slots is pre-determined and slots are filled starting from the highest-score applicant, until exhaustion.

[5] The equality of weights is relevant for this result in combination with the fact that, independently of the Allocation rule, within each site $\tau_j^{ATE} = \tau_j^{ATT} \equiv \lim_{\varepsilon\to 0^+} \mathbb{E}[\tau_{ij}|X^*_{ij} = c_j + \varepsilon, j] = \tau_j^{ATNT} \equiv \lim_{\varepsilon\to 0^+} \mathbb{E}[\tau_{ij}|X^*_{ij} = c_j - \varepsilon, j]$.

Excluding the existence of ties to simplify the exposition,[6] let $i = m_j$ denote the unique marginal participant in site j, i.e., the individual who is allocated the last available slot in this site. Under Allocation rule 2, it is common to set the site-specific cutoff c_j equal to the score of the marginal individual in site j, $X^*_{m_j}$, so that the treatment indicator becomes

$$D_{ij} = \mathbb{I}[X^*_{ij} \geq c_j] = \mathbb{I}[X^*_{ij} \geq X^*_{m_j}]. \tag{12}$$

Therefore, at the end of the allocation process, in each site j there is one program participant who is located exactly at the cutoff c_j, $K_j - 1$ participants who are located above c_j, and $N_j - K_j$ non-participants who are located below c_j. Note that even if the density function of the running variable from which applicants are randomly drawn, $f_{X^*}(x|j)$, is continuous over its entire support (Assumption 3), in the hypothetical experiment of drawing a subject at random from the N_j applicants of site j:

- The probability of drawing the one located exactly at the cutoff c_j is $\frac{1}{N_j}$, where c_j is the empirical $(1 - \frac{K_j}{N_j})$ quantile in the actual sample of N_j applicants (this probability is instead zero under Allocation Rule 1).

- The $N_j - K_j$ unexposed units are random draws from $\frac{f_{X^*}(x|j)}{F_{X^*}(c_j|j)}$; then, the expected number of units in the infinitesimal interval $(x - \varepsilon, x)$, $x < c_j$, is

$$\frac{(N_j - K_j) f_{X^*}(x|j)\varepsilon}{F_{X^*}(c_j|j)}. \tag{13}$$

  Note that $\frac{(N_j - K_j)}{F_{X^*}(c_j|j)}$ differs from N_j only because of sampling variability. As an implication, (13) is approximately equal to $N_j f_{X^*}(x|j)\varepsilon$.

- The expected number of units in the infinitesimal interval $(x, x + \varepsilon)$, $x > c_j$ is

$$\frac{(K_j - 1) f_{X^*}(x|j)\varepsilon}{(1 - F_{X^*}(c_j|j))} \approx N_j \left(1 - \frac{1}{K_j}\right) f_{X^*}(x|j)\varepsilon. \tag{14}$$

Therefore, the expected number of units in the left and right open neighborhoods of the cutoff are, respectively,

$$\frac{(N_j - K_j) f_j \varepsilon}{F_{X^*}(c_j|j)} \approx N_j f_j \varepsilon, \tag{15}$$

[6] We discuss in Section 9 how the analysis changes in the presence of ties.

$$\frac{(K_j - 1) f_j \varepsilon}{(1 - F_{X^*}(c_j|j))} \approx N_j \left(1 - \frac{1}{K_j}\right) f_j \varepsilon, \tag{16}$$

where here and in what follows f_j is shorthand for the density function of the running variable in the population of potential applicants to site j, evaluated at c_j.

Suppose that in this context the researcher adopts NP to estimate τ^{ATE}. To characterize the NP estimand in this setting, consider a third neighborhood of zero that includes 0, denoted $\widetilde{\mathbb{O}}^+$, in addition to those defined in (7) which do not include 0,

$$\widetilde{\mathbb{O}}^+ \;=\; \{X_{ij} : X_{ij} = 0\} \text{ U } \mathbb{O}^+ = (X_{ij} : 0 \leq X_{ij} < +\varepsilon) \quad \text{with} \quad \varepsilon \to 0^+. \tag{17}$$

Under Allocation rule 2, it is this closed neighborhood above the cutoff that is relevant to define the NP estimand, not $\mathbb{O}^+$:

$$\begin{aligned} \tau^{NP_2} &= \mathbb{E}[Y_{ij}|\widetilde{\mathbb{O}}^+] - \mathbb{E}[Y_{ij}|\mathbb{O}^-] \\ &= \sum_{j=1}^{J} \left[\mathbb{E}[Y_{ij}|\widetilde{\mathbb{O}}^+, j]\; P[j|\widetilde{\mathbb{O}}^+]\right] - \sum_{j=1}^{J} \left[\mathbb{E}[Y_{ij}|\mathbb{O}^-, j]\; P[j|\mathbb{O}^-]\right]. \end{aligned} \tag{18}$$

Replacing potential outcomes, estimand τ^{NP_2} coincides with the following population parameter:

$$\begin{aligned} \tau^{NP_2} &= \sum_{j=1}^{J} \left[\mathbb{E}[Y_{ij}(1)|\widetilde{\mathbb{O}}^+, j]\; P[j|\widetilde{\mathbb{O}}^+]\right] - \sum_{j=1}^{J} \left[\mathbb{E}[Y_{ij}(0)|\mathbb{O}^-, j]\; P[j|\mathbb{O}^-]\right] \\ &= \sum_{j=1}^{J} \left[\mathbb{E}[Y_{ij}(1)|0, j]\; P[j|\widetilde{\mathbb{O}}^+]\right] - \sum_{j=1}^{J} \left[\mathbb{E}[Y_{ij}(0)|0, j]\; P[j|\mathbb{O}^-]\right], \end{aligned} \tag{19}$$

where the change in the conditioning term of the expectations in the second line of equation (19) derives from the continuity condition in Assumption 2. Considering Assumption 3 and equations (15) and (16), it is easy to see that, different from τ^{NP_1}, τ^{NP_2} does not coincide with τ^{ATE_1} because

$$P[j|\widetilde{\mathbb{O}}^+] \;=\; \left[\frac{1 + N_j \left(1 - \frac{1}{K_j}\right) f_j \varepsilon}{J + \sum_{j=1}^{J} N_j \left(1 - \frac{1}{K_j}\right) f_j \varepsilon}\right] \neq \frac{1}{J} \frac{N_j f_j}{\overline{(N f)}} = P[j|\mathbb{O}^-], \tag{20}$$

where we make use of the expected number of units in the right and left open neighborhoods of the cutoff in (16) and (15), respectively. In words, the problem originates from the fact that, because of the probability mass at the cutoff under Allocation rule 2, the NP_2 estimand weights the potential outcomes above and below the cutoff with different weights; only the weights below coincide with the ones of the NP_1 estimand in the non-problematic case (which are also the weights that define τ^{ATE_1}).[7] Moreover, note that, under the problematic allocation rule, this probability mass at the cutoff is

$$P[j|0] = \frac{1}{J}. \tag{21}$$

4.1 The "bias" of the NP estimand under Allocation rule 2

To understand the anatomy of the "bias" of τ^{NP_2} relative to a causal parameter, note first that the different probability weights in $\mathbb{O}^-$, 0, and $\mathbb{O}^+$ imply that under Allocation rule 2

$$\tau^{ATE_2} \neq \tau^{ATT_2} \neq \tau^{ATNT_2}. \tag{22}$$

Therefore, the researcher must choose the causal parameter of interest. In what follows we focus the analysis on the "bias" with respect to τ^{ATNT_2}, for two reasons. First, note that under Allocation rule 2,

$$\tau^{ATNT_2} = \tau^{ATE_1}, \tag{23}$$

because both parameters weight the treatment effects in each site with the same weights $P[j|\mathbb{O}^-]$, which are the unique weights of equation (9) under Allocation rule 1 and the weights on the left of the cutoff under Allocation rule 2. A researcher may be interested in assessing if, using the same NP estimation approach, there is a difference between the NP_2 estimand that would be estimated under Allocation rule 2 and the NP_1 estimand that would be estimated under Allocation rule 1. Second, τ^{ATNT_2} is the relevant policy parameter if the researcher is interested in the consequences of marginally expanding the number of slots in the different sites.[8]

[7] In any multi-site setting, a similar difference in weights on the two sides of the cutoff would arise even under Allocation rule 1, if Assumption 3 is violated at some cutoff c_j while Assumptions 1 and 2 hold.

[8] In any case, the discussion that follows can be easily modified to study the bias of the NP_2 estimand with respect to the ATE_2 or the ATT_2, if a reader were interested in these alternatives.

In light of this observation, the NP estimand can be decomposed as follows,

$$\begin{aligned}
\tau^{NP_2} &= \mathbb{E}[Y_{ij}|\widetilde{\mathbb{O}}^{+}] - \mathbb{E}[Y_{ij}|\mathbb{O}^{-}] \\
&= \sum_{j=1}^{J} \mathbb{E}[Y_{ij}(1) - Y_{ij}(0)|0,j]\; P[j|\mathbb{O}^{-}] + \sum_{j=1}^{J} \mathbb{E}[Y_{ij}(1)|0,j]\; \left(P[j|\widetilde{\mathbb{O}}^{+}] - P[j|\mathbb{O}^{-}]\right) \\
&= \tau^{ATNT_2} + \sum_{j=1}^{J} \mathbb{E}[Y_{ij}(1)|0,j]\; \left(P[j|\widetilde{\mathbb{O}}^{+}] - P[j|\mathbb{O}^{-}]\right). \qquad (24)
\end{aligned}$$

The "bias" in the last line of (24) is not zero, in general. It is equal to zero under Allocation rule 1 because in that case $P[j|\widetilde{\mathbb{O}}^{+}] = P[j|\mathbb{O}^{+}] = P[j|\mathbb{O}^{-}]$.[9] This "bias" can be further decomposed using the fact that

$$P[j|\widetilde{\mathbb{O}}^{+}] = P[j|0]P[0|\widetilde{\mathbb{O}}^{+}] + P[j|\mathbb{O}^{+}]P[\mathbb{O}^{+}|\widetilde{\mathbb{O}}^{+}], \qquad (25)$$

since it follows from (17) that $P[0|\widetilde{\mathbb{O}}^{+}] + P[\mathbb{O}^{+}|\widetilde{\mathbb{O}}^{+}] = 1$, which yields the following expression for the NP_2 estimand:

$$\begin{aligned}
\tau^{NP_2} &= \tau^{ATNT_2} \\
&+ \underbrace{\sum_{j=1}^{J} \mathbb{E}[Y_{ij}(1)|0,j]\; \left(P[j|0] - P[j|\mathbb{O}^{-}]\right)}_{B_1} P[0|\widetilde{\mathbb{O}}^{+}] \\
&+ \underbrace{\sum_{j=1}^{J} \mathbb{E}[Y_{ij}(1)|0,j]\; \left(P[j|\mathbb{O}^{+}] - P[j|\mathbb{O}^{-}]\right)}_{B_2} P[\mathbb{O}^{+}|\widetilde{\mathbb{O}}^{+}]. \qquad (26)
\end{aligned}$$

We next analyze the three components of the "bias", i.e., B_1, B_2, and $P[0|\widetilde{\mathbb{O}}^{+}]$. Given the definitions of $P[j|\mathbb{O}^{-}]$ and $P[j|0]$ in equations (20) and (21), component B_1 is (the negative of) the covariance across sites between the average outcome of each site in the presence of

[9] Equation (24) shows also why the analysis of a Sharp NP design that we carry out here is sufficient to understand the problems a researcher would face in a Fuzzy NP design, which we do not discuss for brevity. The Fuzzy RD estimand is the ratio of two Sharp RD estimands: the ITT effect of the assignment on the outcome at the numerator and the ITT effect of the assignment on the treatment at the denominator. The "bias" in the numerator is the same as in equation (24); the "bias" in the denominator is analogous to the one in the numerator, with $Y(1)$ replaced by $D(1)$. In general, there is no reason for these two biases to cancel out.

treatment and the weight of the site in the neighborhood below the cutoff:[10]

$$B_1 = -\frac{1}{J}\sum_{j=1}^{J}\mathbb{E}[Y_{ij}(1)|0,j]\left(\frac{N_j f_j}{\overline{(Nf)}} - 1\right) = -\text{Cov}\left[\mathbb{E}[Y_{ij}(1)|0,j], \frac{N_j f_j}{\overline{(Nf)}}\right]. \tag{27}$$

For example, in an educational context, the absolute value of this covariance is large when schools that attract students with the best outcomes in case of treatment are also more popular, i.e., they attract a larger fraction of applicants in general and in a neighborhood of the cutoff. Note that bias component B_1 depends on the sites' weights, but given those weights B_1 is independent of total sample size N.

In equation (26), component B_1 is multiplied by the second component of the "bias",

$$P[0|\widetilde{\mathbb{O}}^+] = \frac{J}{J + \sum_{j=1}^{J} N_j\left(1 - \frac{1}{K_j}\right) f_j \varepsilon} = \frac{1}{1 + \varepsilon\left(\overline{(Nf)} - \overline{h}\right)}, \tag{28}$$

where $\overline{h}$ is the cross-site average of $h_j = \frac{N_j f_j}{K_j}$, which is the hazard function of the running variable evaluated at the cutoff. This second component is the weight in $\widetilde{\mathbb{O}}^+$ of the marginal treated subjects located exactly at the cutoff. Such weight would be equal to zero under Allocation rule 1 (because in that case the probability of observing subjects located exactly at the cutoff is zero) and so the "bias" component B_1 would become irrelevant, even if large, in the overall NP "bias". Under Allocation rule 2 instead, the NP approach leads to a "bias" because $P[0|\widetilde{\mathbb{O}}^+]$ is not zero – unless the number of applicants per site increases to infinity, as is evident in equation (28) – and so the B_1 component becomes relevant.

Finally, the last component can be written as

$$\begin{aligned} B_2 &= \frac{\overline{h}}{\overline{(Nf)} - \overline{h}}\left\{\text{Cov}\left[\mathbb{E}[Y_{ij}(1)|0,j], \frac{N_j f_j}{\overline{(Nf)}}\right] - \text{Cov}\left[\mathbb{E}[Y_{ij}(1)|0,j], \frac{h_j}{\overline{h}}\right]\right\} \\ &= \frac{\overline{h}}{\overline{(Nf)} - \overline{h}}\left\{-B_1 - B_3\right\}, \qquad (29) \end{aligned}$$

where the first term of the difference inside the curly brackets is again (minus) B_1, while the second, B_3, is the covariance between the average outcome in a site in the presence of treatment and the relative hazard function of the running variable evaluated at the cutoff.

[10] Strictly speaking the weight of a site is $\frac{N_j f_j}{J\overline{(Nf)}}$, while $\frac{N_j f_j}{\overline{(Nf)}}$ is proportional to the weight.

"Bias" component B_2 vanishes when the number of applicants goes to infinity, and is probably (but not necessarily) small whatever the number of applicants because the ratio $\frac{\overline{h}}{\overline{(Nf)}-\overline{h}}$ is bound to be small in practice. The reason is that in every site the number of applicants in a neighborhood of the cutoff $(N_j f_j)$ is small relative to the number of slots in the entire site (K_j); therefore, $\overline{h}$ is a small number relative to $\overline{(Nf)}$. Note, however, that since the sign of B_3 depends on the shape of the cumulative density function of the running variable, we cannot rule out that $(-B_1 - B_3)$ is large in some applications, and therefore a general conclusion about the magnitude of this component cannot be drawn.[11]

It is important to realize that a *necessary* condition for the existence of a "bias" of the NP estimator under Allocation rule 2 is that the J sites are heterogeneous with respect to both the average potential outcome $Y(1)$ at the cutoff and the number of applicants per site. In particular, the role played by the variance across sites of the relative site size $\left(\frac{N_j f_j}{\overline{(Nf)}}\right)$ is clear from equation (27). This variance *decreases* – and so the NP "bias" also decreases – as the number of applicants in each site shrinks, keeping the total number of sites fixed at J (see the Online Appendix to this section for details).

Summing up, under Allocation rule 2:

- A *necessary* condition for the "bias" of NP with respect to the $ATNT_2$ is that the J sites are heterogeneus with respect to the average potential outcome at the cutoff as well as with respect to the number of applicants per site.

- If this necessary condition is met then the size of the "bias" is mainly determined by the covariance $-B_1$ across sites between the outcome of applicants in the presence of treatment and their relative number, in a neighborhood of the cutoff.

- Covariance $-B_1$ enters the "bias" weighted by $P[0|\widetilde{\mathbb{O}}^+]$, the proportion of subjects located at the cutoff out of those in a left-closed neighborhood above the cutoff. This weight goes to zero as the number of applicants per site diverges to infinity; this is one reason why the aforementioned condition is only necessary for a "bias".

[11] We can establish that B_2 becomes zero in the very special case in which all sites offer exactly the same number of slots. In this case, $K_j = K$ for all j and the relative popularity coincides with the relative degree of rationing, thus the two covariances in equation (29) are equal and B_2 is zero.

- The size of the additional "bias" component B_2 is uncertain. It may be small because it depends on the ratio between the number of applicants in a neighborhood of the cutoff and the number of slots in the entire site, which is typically small; but we cannot rule out that it is large in some applications, because its other component, $(-B_1 - B_3)$, can be sizeable.

- Altogether, the "bias" of NP with respect to the $ATNT_2$ is likely to be small when the number of applicants per site is large – in which case weight $P[0|\widetilde{\mathbb{O}}^+]$ is negligible – or when the number of applicants is very small in each site – in which case it is the variability of the weights across sites that is negligible. In all other cases, the "bias" may be large.

4.2 Does it help to drop observations located at the cutoff?

In light of the analysis carried out so far, it is natural to ask whether dropping the observations located exactly at the cutoff would eliminate the "bias" of the NP estimand under Allocation rule 2. The answer is: not necessarily. To see why, note that after removing these marginal subjects the NP estimand is given by

$$\begin{aligned}\tau^{NP_3} &= \mathbb{E}[Y_{ij}|\mathbb{O}^+] - \mathbb{E}[Y_{ij}|\mathbb{O}^-] \\ &= \sum_{j=1}^{J}\left[\mathbb{E}[Y_{ij}|\mathbb{O}^+, j]\ P[j|\mathbb{O}^+]\right] - \sum_{j=1}^{J}\left[\mathbb{E}[Y_{ij}|\mathbb{O}^-, j]\ P[j|\mathbb{O}^-]\right].\end{aligned} \tag{30}$$

This expression looks identical to equation (8), which defines τ^{NP_1}. However, here the weights used to derive the average potential outcomes across sites above and below the zero cutoff do *not* coincide, contrary to equation (8). In particular, $P[j|\mathbb{O}^+]$ is not the weight that we would use under Allocation rule 1. The reason is that if the marginal subjects located at each cutoff are removed, the expected number of subjects in an open neighborhood above the zero cutoff is as in equation (16), and so

$$P[j|\mathbb{O}^+] = \frac{1}{J}\frac{N_j\left(1-\frac{1}{K_j}\right)f_j}{\overline{Nf}-\frac{\overline{Nf}}{K}} \neq \frac{1}{J}\frac{N_j f_j}{\overline{Nf}} = P[j|\mathbb{O}^-]. \tag{31}$$

As a result of the difference between the two weights, we can write

$$\begin{aligned}\tau^{NP_3} &= \tau^{ATNT} + \sum_{j=1}^{J} \mathbb{E}[Y_{ij}(1)|0,j] \ \left(P[j|\mathbb{O}^{+}] - P[j|\mathbb{O}^{-}]\right) \\ &= \tau^{ATNT} + B_2,\end{aligned} \tag{32}$$

which indicates that under Allocation rule 2 the strategy of dropping the subjects located exactly at the cutoff would still induce the B_2 bias component defined in equation (29).

For the reasons discussed in Section 4.1, B_2 is likely to be small when the number of applicants per site diverges to infinity as well as when the number of applicants is very small in each site.[12] But we cannot exclude that it is large in some settings. It is then an empirical question whether the strategy of dropping the marginal subjects who are located at the cutoffs is enough to eliminate any relevant "bias" or not.

5 Fixed effects: the safest option

The previous section has established that under Allocation rule 2, the NP estimand – whether including or excluding marginal subjects – in general does not coincide with a meaningful causal parameter. In this section we illustrate a Site Fixed Effect (SFE) estimation strategy that eliminates any "bias" resulting from the NP strategy.

5.1 The fixed effect estimator for the multi-cutoff RDD

Consider the following SFE estimator,

$$\begin{aligned}\hat{\tau}^{SFE} &= \frac{\sum_{j=1}^{J}\sum_{i=1}^{N_j} R_{ij}\left(Y_{ij} - Y_{.j}\right)\left(D_{ij} - D_{.j}\right)}{\sum_{j=1}^{J}\sum_{i=1}^{N_j} R_{ij}\left(D_{ij} - D_{.j}\right)^2} \\ &= \sum_{j=1}^{J} \omega_j \frac{\sum_{i=1}^{N_j} R_{ij}\left(Y_{ij} - Y_{.j}\right)\left(D_{ij} - D_{.j}\right)}{\sum_{i=1}^{N_j} R_{ij}\left(D_{ij} - D_{.j}\right)^2},\end{aligned} \tag{33}$$

where R_{ij} is a kernel function appropriately designed to include only observations in a neighborhood of the cutoff and $Y_{.j}$ denotes the local average outcome in site j evaluated using

[12] As explained in Section 4.1, it is also small in some special cases: when $K_j = K$ for all j and when K_j is large in each site.

the same kernel function R_{ij}, and similarly for the treatment indicator $D_{.j}$.[13] The weight of each site is given by

$$\omega_j = \frac{\sum_{i=1}^{N_j} R_{ij}\left(D_{ij} - D_{.j}\right)^2}{\sum_{j=1}^{J}\sum_{i=1}^{N_j} R_{ij}\left(D_{ij} - D_{.j}\right)^2}. \tag{34}$$

Thus, $\hat{\tau}^{SFE}$ is the weighted average of the local (kernel-weighted) site-specific regression coefficients of the observed outcome on the treatment exposure indicator. Only sites with at least one treated and one untreated unit within a neighborhood of the cutoff receive a strictly positive weight. If treatment exposure is assigned according to Allocation rule 1, after replacing potential outcomes the estimand of this estimator coincides with

$$\begin{aligned}
\tau^{SFE} &= \sum_{j=1}^{J} \omega_j^{SFE}\left(\mathbb{E}[Y_{ij}(1)|\mathbb{O}^+, j] - \mathbb{E}[Y_{ij}(0)|\mathbb{O}^-, j]\right) \\
&= \sum_{j=1}^{J} \omega_j^{SFE}\left(\mathbb{E}[Y_{ij}(1)|0, j] - \mathbb{E}[Y_{ij}(0)|0, j]\right) \qquad (35)
\end{aligned}$$

which is a weighted average of the site-specific Average Treatment Effects at the cutoffs.[14] As noted above, independently of the allocation rules, SFE drops sites with units only on one side of the cutoff within the neighborhood, and thus uses only sites for which a proper counterfactual comparison is possible. In this respect, NP does not offer a better solution, because it uses all sites at the cost of including also those without a counterfactual comparison, for which a treatment effect cannot be identified.

Moreover, under Allocation rule 1, owing to the absence of subjects located exactly at the cutoff, the weights coincide with those of the non problematic case (equation 9), i.e.,

$$\omega_j^{SFE} = \frac{1}{J}\frac{N_j f_j}{\overline{(Nf)}} = P[j|\mathbb{O}^+] = P[j|\mathbb{O}^-] = P[j|0], \tag{36}$$

so in this case each site's weight is proportional to the number of subjects in a neighborhood of the cutoff. Thus, under Allocation rule 1, τ^{SFE} coincides with the ATE across all sites.

Under Allocation rule 2, instead, the weights of the J sites are no longer proportional to $N_j f_j$. The SFE estimand assigns a strictly positive weight only to sites with at least one

[13] See Section 8 for details on how we select the bandwidth to implement this estimator in our empirical analysis.

[14] Once again, the change in the conditioning term of the expectations in the last line of (35) derives from the continuity condition in Assumption 2.

treated and one untreated unit. Therefore, in the j^{th} site the expected number of units in a (open) left and (closed) right neighborhood of the cutoff are, respectively,

$$1 + N_j f_j \epsilon (1 - \frac{N_j - K_j - 1}{N_j - K_j}); \tag{37}$$

$$1 + N_j f_j \epsilon (1 - \frac{K_j - 1}{K_j}). \tag{38}$$

Using these expected numbers of units, the probability of exposure in a neighborhood of the cutoff in site j is

$$0.5 + \frac{1}{(f_j \epsilon)^{-1} + N_j - 2}(r_j - 1), \tag{39}$$

where $0 < r_j = \frac{2K_j}{N_j} < 2$, so that when the number of available slots is equal to half the number of applicants the probability of exposure (at the cutoff) is 0.5.[15]

Note however that even when $r_j \neq 1$, the deviation of this probability from 0.5 is likely to be negligible in practice because (a) the numerator is below 1 in absolute value; (b) the denominator is the sum of the number of applicants to site j (minus 2) plus the inverse of the probability to have a unit in the neighborhood of the cutoff, which is a "small" probability.

In conclusion, even under Allocation rule 2, the probability (at the cutoff) of being treated is close to 0.5 in each site and the weight of each site is approximately proportional to $N_j f_j$, as under Allocation rule 1. Therefore, even under Allocation rule 2 the estimand of SFE given by equation (35) coincides approximately with the ATE across sites.

5.2 Efficiency of the SFE estimator in multi-cutoff RD

The ability to identify a meaningful causal parameter is not the only advantage of the SFE estimator in a multi-cutoff RD design. In such a setting, this estimator is likely more precise than the NP estimator. To see why, consider the following linear specification for the effect on outcome Y_{ij} of the exposure of subject i to treatment D_{ij} in site j:

$$Y_{ij} = \tau D_{ij} + u_j + e_{ij}. \tag{40}$$

[15] Eq. (39) is a first order Taylor expansion of the probability of exposure as a function of r_j around $r_j = 1$.

As in previous sections, this equation is intended to be estimated on the sub-sample of subjects in a suitable interval around the cutoff. The site fixed-effect u_j represents the site-specific average outcome (at the cutoff) in case $D_{ij} = 0$, while e_{ij} represents the unit-specific unobservable component of the outcome. The sampling variance of the SFE estimator is

$$\mathrm{Var}[\hat{\tau}^{SFE}] = \frac{\mathrm{Var}[e]}{SS_w}, \tag{41}$$

where SS_w is the within-site sum of squared deviations of the treatment indicator D_{ij} from its mean. The corresponding sampling variance of the NP estimator is:

$$\mathrm{Var}[\hat{\tau}^{NP}] = M\frac{\mathrm{Var}[u] + \mathrm{Var}[e]}{SS_w + SS_b} \tag{42}$$

where SS_b is the between-site sum of squared deviations of the treatment indicator D_{ij} from its mean and M is the Moulton Factor (see Angrist and Pischke, 2008, Section 8.2.1).

Since $M \geq 1$, a *sufficient* condition for the SFE estimator to be more precise than the NP estimator is that

$$\frac{\mathrm{Var}[u]}{\mathrm{Var}[u] + \mathrm{Var}[e]} > \frac{SS_b}{SS_b + SS_w}, \tag{43}$$

where the LHS is the intraclass correlation of the composite error term $u_j + e_{ij}$,[16] while the RHS is the intraclass correlation of the treatment indicator D_{ij}. In general, one cannot tell whether condition (43) is satisfied or not because, comparing the right-hand sides of (41) and (42), one cannot tell whether the gain induced by dropping $\mathrm{Var}[u]$ in the numerator is larger than the loss from dropping SS_b in the denominator. However, in the RD setting that we study this condition is likely to be met in most applications for the following reason.

As we show in the Online Appendix, under Allocation rule 1 the intraclass correlation of the treatment indicator on the RHS of (43) is:

$$\frac{SS_b}{SS_b + SS_w} = \left(1 - 2\overline{\left(\frac{1}{n}\right)}\right)\frac{1}{\bar{n}}, \tag{44}$$

where $\bar{n}$ is the average of n_j, the number of subjects per site in a neighborhood of the cutoff, and $\overline{\left(\frac{1}{n}\right)}$ is the average of $\frac{1}{n_j}$. Since, using SFE, $n_j \geq 2$, in the extreme case in which $n_j = 2$

[16] Here and in the rest of the paper, the correlations on the two sides of (43) are labelled as “intraclass” in line with the literature, even if in our context a class is a site.

in all sites – and if $\bar{n}$ diverges to infinity – the RHS of (43) is zero. Therefore, even a tiny value of $\mathrm{Var}[u]$ is enough to guarantee an efficiency advantage of SFE relative to NP. These are the cases in which a RD design coincides (locally at the cutoff) with a stratified RCT with an equal proportion of treated in each stratum (see Athey and Imbens, 2017). In this case, there is no between-site variation of the treatment indicator, and so conditioning on site fixed effects generates *only* an efficiency gain – this conditioning reduces the variance of the error term without any countervailing effect.

When n_j varies across sites, the RD setting deviates from the case of a stratified RCT with an equal proportion of treated in each stratum, because of sampling variability. However,[17]

$$\left(1 - 2\overline{\left(\frac{1}{n}\right)}\right)\frac{1}{\bar{n}} \leq \left(1 - \frac{2}{\bar{n}}\right)\frac{1}{\bar{n}}, \tag{45}$$

and so a sufficient condition for SFE to be more precise than NP is

$$\frac{\mathrm{Var}[u]}{\mathrm{Var}[u] + \mathrm{Var}[e]} > \left(1 - \frac{2}{\bar{n}}\right)\frac{1}{\bar{n}}. \tag{46}$$

The RHS of (46) reaches the maximum at $\bar{n} = 4$, and the maximum is as large as 0.125. Therefore, even in this case a moderate degree of unobservable heterogeneity across sites would be enough to confer the SFE estimator an efficiency advantage.[18]

Under Allocation rule 2, the deviation from the case of a stratified RCT with an equal proportion of treated in each stratum is of course larger because of the presence of a subject at each cutoff in a context in which the sample size in a neighborhood of the cutoff is generally different across sites. Still, using the same argument used in Section 5.1 in relation to equation (39), the deviation from 0.5 of the probability of exposure due to Allocation rule 2 is plausibly negligible in practice. Therefore, SFE has an efficiency advantage also under Allocation rule 2, for the same reasons it has one under Allocation rule 1.

[17] Recall that the armonic average is smaller than or equal to the arithmetic average.

[18] To simplify the exposition, these calculations are derived as if the NP and SFE are evaluated on the same set of sites. In general, this is not the case because, as noted above, SFE assigns a strictly positive weight only to sites with positive variance of the treatment status while NP also uses sites where this variance is zero. However, when NP uses more sites than SFE, the value of SS_w is exactly the same for the two estimators. Any difference in the size of the two sets of sites would affect only the value of SS_b. Nonetheless, our argument for using the sufficient condition (46) is that a fair comparison between NP and SFE should condition on the set of sites within which both treated and untreated units are available (around the cutoff). Unless one has strong reasons to think that site membership is not a confounder.

6 Alternative ways of defining the cutoff

There are alternative strategies based on re-defining the cutoff in a way that sidesteps the presence of an individual located exactly at each cutoff under Allocation Rule 2. This section shows that these strategies do not necessarily solve the identification problem and, in addition, they are likely to be less precise than SFE (for the same reasons discussed in the previous section in relation to NP).

6.1 Symmetric distance Normalizing-and-Pooling

The first solution, which we label as "SYM", is commonly used in political sciences, where a site is an election. For example, Boas and Hidalgo (2011) and Kendall and Rekkas (2012) introduce an asymmetry between the last subject who is exposed to the treatment – the marginally treated – and the first subject who is not – the marginally non-treated. These authors posit that, in each site j, the cutoff that is relevant for the treated (resp. non-treated) is the score of the marginally non-treated (resp. treated). Formally, let $X^*_{m_j}$ and $X^*_{d_j}$ denote the scores of, respectively, the marginally treated and marginally non-treated ("denied") subjects. The normalized scores are then computed as $X_{ij} = X^*_{ij} - X^*_{d_j}$ for each unit i who is exposed to the treatment, and $X_{ij} = X^*_{ij} - X^*_{m_j}$ for each unit i who is not exposed to the treatment. As a consequence, there is no unit located at any cutoff, i.e., there is no i for whom $X_{ij} = 0$. After defining the running variable this way, the authors proceed pooling together the J sites exactly as in NP.

In finite samples, this way of redefining the normalized score induces a discontinuity in the regression of $Y_{ij}(0)$ on this score at the normalized cutoff, which is given by

$$\begin{aligned} & \mathbb{E}[Y_{ij}(0)|X_{ij} = \mathbb{O}^+, j] - \mathbb{E}[Y_{ij}(0)|X_{ij} = \mathbb{O}^-, j] \\ = \ & \mathbb{E}[Y_{ij}(0)|X^*_{ij} = X^*_{d_j}, j] - \mathbb{E}[Y_{ij}(0)|X^*_{ij} = X^*_{m_j}, j]. \end{aligned} \quad (47)$$

This difference is *not* zero as long as the regression of $Y(0)$ on X^* is not flat around the cutoff(s). As the number of individuals per site, N_j, grows to infinity the difference on the RHS of equation (47) goes to zero. This is so because in large samples, $X^*_{m_j} - X^*_{d_j}$ converges to zero. How much it matters in finite samples is an issue that we explore in Section 8.

6.2 Split-the-distance Normalizing-and-Pooling

An alternative way of avoiding one subject located exactly at each cutoff is the following. Consider again the scores of the marginally treated and marginally non-treated ("denied") subjects, $X^*_{m_j}$ and $X^*_{d_j}$. One can define the cutoff in site j as a weighted average of the two,

$$\alpha X^*_{m_j} + (1-\alpha)X^*_{d_j}, \tag{48}$$

where $\alpha \in (0,1)$. The idea is that *any* score above $X^*_{d_j}$, even if strictly below $X^*_{m_j}$, would ensure treatment to subject m_j. The researcher can then normalize the score in each site using this average cutoff and pool across sites in the usual way. We label this strategy as "SPLIT".

By definition, in each site there is one treated unit located exactly at a $(1-\alpha)(X^*_{m_j} - X^*_{d_j})$ distance on the right of the cutoff and one untreated unit located exactly at a $\alpha(X^*_{m_j} - X^*_{d_j})$ distance on the left of the cutoff. Whether this strategy works or not depends again on the weight of each site on the left and on the right of the cutoff. Setting $\alpha = 0.5$, the expected number of units in a finite right neighborhood of the cutoff is approximately equal to

$$1 + N_j \left(1 - \frac{1}{K_j}\right) f_j \nu, \tag{49}$$

where ν is the width of the neighborhood. In a finite left neighborhood, the corresponding expected number is

$$1 + N_j \left(1 - \frac{1}{N_j - K_j}\right) f_j \nu. \tag{50}$$

Depending on the value of the ratio $\frac{K_j}{N_j}$, these two numbers – and so the weights of the site on the left and on the right of the cutoff – will generally differ. In this case, following the discussion of the NP bias in Section 4.1, the resulting estimand of the causal effect would be biased to a degree depending on the cross-site covariance between $\mathbb{E}[Y(1)|0,j]$ and $(P[j|\mathbb{O}^+] - P[j|\mathbb{O}^-])$. We explore empirically the performance of this estimator in Section 8, at different levels of the number of applicants per site.

7 Rank-Distance: a strategy that should be avoided

Another possible strategy in the multi-cutoff RD setting under Allocation rule 2 consists of following the NP approach after replacing the original normalized running variable with a normalized running variable in terms of ranks (e.g., Abdulkadiroglu et al., 2014). This solution comes in two flavors. One consists of first creating the running variable site by site in terms of ranks and then normalizing and pooling the observations with respect to the running variable in ranks (RKNP). The other inverts the order of these operations (NPRK). In this section we show that these strategies are either potentially problematic or superfluous.

Starting with RKNP, let $x_{(1)}$,$x_{(2)}$, …, $x_{(N_j-K_j)}$, , $x_{(N_j-K_j+1)}$, , …$x_{(N_j)}$ be the list of scores ordered from the smallest to the largest in site j. $N_j - K_j$ individuals (i.e. those in the first $N_j - K_j$ positions of this list from the bottom) are not offered a slot in site j, the remaining individuals, in positions between $N_j - K_j + 1$ and N_j, are instead exposed to the treatment in the same site. The normalized (and standardized) running variable in rank is:

$$\rho^*_{ij} = 100\frac{\rho_{ij} - N_j + K_j - 1}{N_j}, \tag{51}$$

where ρ_{ij} is the rank of the i^{th} applicant in site j.

In each site, there will be one marginal subject with $\rho^*_{ij} = 0$, while in any two symmetric open neighborhoods of the cutoff of size ε, which we denote again $\mathbb{O}^- = (-\varepsilon, 0)$ and $\mathbb{O}^+ = (0, +\varepsilon)$, the number of subjects will be:

$$n_j = \begin{cases} 0 & \text{if} \quad N_j\varepsilon < 1 \\ 1 & \text{if} \quad 1 < N_j\varepsilon < 2 \\ 2 & \text{if} \quad 2 < N_j\varepsilon < 3 \\ \dots & \end{cases} \tag{52}$$

That is, n_j is proportional to N_j. Defining in terms of the normalized ranking distance also the closed neighborhood $\widetilde{\mathbb{O}}^+ = \{0\} \text{ U } \mathbb{O}^+ = [0, +\varepsilon)$, the corresponding weight of site j in this neighborhood is

$$P[j|\widetilde{\mathbb{O}}^+] = \frac{1}{J}\frac{1 + n_j}{1 + \bar{n}}, \tag{53}$$

where $\bar{n}$ is the average of n_j across sites, while in $\mathbb{O}^-$ the weight is

$$P[j|\mathbb{O}^-] = \frac{1}{J}\frac{n_j}{\bar{n}}. \tag{54}$$

The fact that the two weights in (53) and (54) differ leads to the same conclusions reached in Section 4.1: the RKNP estimator that includes the subject located exactly at the cutoff in general does not identify a meaningful causal parameter under Allocation rule 2, even if the RD design is valid in each site. However, removing subjects located at the cutoff when using the normalized rank-distance as the running variable restores identification, because in this case both sets of weights, above and below are equal and given by

$$P[j|\mathbb{O}^-] = \frac{1}{J}\frac{n_j}{\bar{n}} = P[j|\mathbb{O}^+]. \tag{55}$$

This is promising but is not enough to solve the problem, for a different reason. The weights in equation (55) are proportional to N_j, which is the total number of applicants to site j, and, because of (52), yield the following estimand,

$$\begin{aligned}
\tau^{RKNP} &= \sum_{j=1}^{J}\left[\mathbb{E}[Y_{ij}|\mathbb{O}^+, j]\,\frac{N_j}{N}\right] - \sum_{j=1}^{J}\left[\mathbb{E}[Y_{ij}|\mathbb{O}^-, j]\,\frac{N_j}{N}\right] \\
&= \sum_{j=1}^{J}\frac{N_j}{N}\left[\mathbb{E}[Y_{ij}(1)|\mathbb{O}^+, j] - [\mathbb{E}[Y_{ij}(0)|\mathbb{O}^-, j]\right].
\end{aligned} \tag{56}$$

However, the "natural" weight that a site should receive in a pooled RD design should depend on the number of applicants located only in a neighborhood of the cutoff, i.e. $N_j f_j$, not on all the applicants to that site, i.e. N_j. Consider for example two sites, j and j', with the same number of applicants, $N_j = N_{j'}$. Suppose that the distribution of running variable is bell-shaped in both sites, but in site j the cutoff is located near the mode while in site j' it is located in one of the tails. Using the RKNP estimation strategy, the two sites would receive the same weight while the RD logic requires that site j should receive a larger weight. This problem would not arise only in very particular and unlikely cases in which both the distributions of the running variable and the cutoffs are the same in the two sites (i.e., $f_j = f_j' = f$ and $c_j = c_j' = c$). Outside of these special cases, the omission of f_j from the weights in (54) would generally make the estimand τ^{RKNP} difficult to interpret.

This estimand would still be a weighted difference between the expected potential outcomes with and without treatment, but it could be "biased" with respect to the standard causal parameters in which researchers are typically interested, as we show in Section 8.

This problem could be handled by re-weighting sites using (an estimate of) f_j. However, in the presence of relevant heterogeneity across sites in the number of applicants, the way in which neighborhoods are defined in equation (55) results in an estimator that would would assign zero weight to sites with a small number of applicants, regardless of re-weighting.

As for the NPRK approach, it is easy to see that under Allocation rule 2 this strategy faces exactly the same problems as the standard NP strategy analyzed in Section 4.1. The reason is that this procedure implies a one-to-one mapping between the original scale of the normalized running variable and the $(0, 100)$ scale of the normalized running variables in ranks. As a result, the weight of each site in the original scale is identical to its weight in the rank scale and would differ on the two sides of the cutoff, as in equation (20). The resulting estimand τ^{NPRK} would be "biased" for the ATT, like τ^{NP_2} in equation (26). Similarly, the properties of the NPRK estimator after removing marginal subjects would be the same as in the NP case without a transformed running variable. For this reason, we do not analyze this estimator in detail and we do not consider it in the empirical illustration that follows.

8 An empirical illustration

We illustrate empirically the results of our econometric analysis using the data of Pop-Eleches and Urquiola (2013). These authors – referred to as PEU in what follows – employ a multi-cutoff RD design to study the effects on students' outcomes of being admitted, at the end of middle school, to oversubscribed "higher-achievement" high schools in Romania. The PEU dataset is ideal for our purposes because rationing is resolved using Allocation rule 2 described in Section 4, which we have shown to be potentially problematic. The running variable that determines admission to one of these high schools is defined by PEU as a "transition score". For each applicant, this is equal to the average between a standardized nationwide test score and middle school GPA. The primary focus in PEU is the effect of the quality of a high school, as measured by the average transition score of its students, on the participation and the grade obtained in a high-stake Baccalaureate exam that determines access to tertiary education.

For the purpose of our empirical illustration, we focus on the sharp RD design in PEU's analysis and so on the Intention To Treat effect of being admitted to these higher-achievement educational institutions on the Baccalaurate grade.[19] After applying our selection criteria, we study a setting that features 1,729 sites (higher-achievement high schools) and 202,424 applicants (in 2001-2003). Since a student can apply to more than one high school, our estimation exercise can leverage about 1.24 million applications, with an average of 716 applicants per site. Thus, we can compare the performance of the estimators described in the previous sections at the original, very large, sample size and when this size is progressively reduced, reaching levels that are closer to those that prevail in the literature. This is another important reason why the PEU dataset is ideal to illustrate our results.[20] Following PEU, in our empirical analysis we take into account multiple applications by clustering standard errors at the individual level.[21]

8.1 Setting-up the PEU sample for the comparison exercise

Referring the interested reader to PEU for all the institutional details, in this section we describe the specific sub-sample of their data that we use for our empirical illustration.

We start from the initial administrative PEU sample of students in the 2001-2003 cohorts, and we follow these authors in restricting the analysis to applicants with a running variable within 1 point from a school admission cutoff. This data set has 1,869,709 observations for 1984 sites. We then drop the observations with missing baccalaureate grade (see footnote 19), those in sites without rationing, and those that have only one applicant to the right

[19] PEU show that about two thirds of Romanian high-school graduates take the Baccalaureate exam, but this fraction is independent of whether a student is admitted or not to "higher achievement" high schools, based on the transition score. Thus, taking the Baccalaureate exam is orthogonal to our ITT indicator. We therefore follow PEU in not worrying about a potential selection problem related to the possibility that a Baccalaureate grade exists only for students who take the corresponding exam.

[20] The PEU data set is used also by Bertanha (2020) to discuss extrapolation in RD settings.

[21] While our theoretical analysis in Sections 3–7 assumes that all applicants are attached to one and only one site, in PEU each subject can apply to multiple sites. However, according to PEU, "regressions restricted to students in bands close to the cutoffs in fact rarely use student-level observations more than once"' (p. 1302), i.e., the case they consider does not depart in a relevant way from our set-up. Nonetheless, we follow PEU in correcting standard errors for for the correlation between the multiple applications of each individual.

or to the left of the cutoff. This cleaning leaves us with a usable sample of 1,729 sites and 1,238,418 applicants. We will refer to this as the *Full Sample*.

This data set contains "ties" within each site, i.e., applicants with the same value of the running variable. To simplify the comparison of the estimators that we consider without having to worry about ties (in particular those located exactly at a cutoff), we add a tiny noise to the running variable, site by site and separately on the two sides of each cutoff, making sure that this noise is sufficiently small to preserve the treatment status of each applicant. This alteration of the original data is irrelevant for our findings (more on this in Section 9).

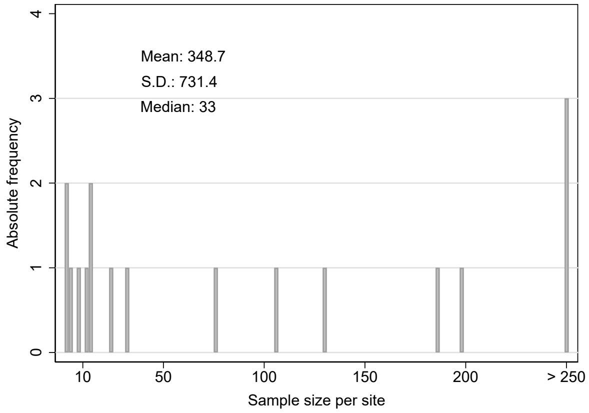
In light of these changes, it is clear that the Full Sample is different from all the samples used in PEU. Therefore, none of our results is directly comparable with the results obtained by these authors (although, for completeness, we report below the PEU estimate that is more comparable to ours). This is not a problem because we are not interested in validating their results, which are perfectly fine: our goal is just to use their data in the best way to illustrate our analysis.

8.2 Sample size reduction

Figure 1 shows the sample size distribution of the data sets used in the 21 studies reviewed in Table A–1 , which are all characterized by multiple cutoffs under Allocation rule 2. It is evident that these sample sizes are in general considerably smaller than the one in PEU. Since our goal is to compare the estimators described in previous sections not only at the very large PEU sample size but also at the smaller sample sizes that are typical in the literature, we need a rule to progressively drop observations from the PEU sample.

There are many ways to implement this sample reduction. The theoretical results described in previous sections have been obtained *for a given number of sites*, and so in order to ensure that the empirical illustration matches our theoretical results, we need a way to progressively drop observations without changing the number of sites. This requirement poses a problem because the support of the distribution of applicants per site in the Full Sample ranges between 10 and 2,224; if we dropped observations proportionally in each site, some sites would disappear in the process. We handle this problem as follows. Starting from the

Figure 1: Distribution of sample size per site across the studies listed in Table A–1



Notes: Considering the 19 studies, different from Pop-Eleches and Urquiola (2013), that are listed in Table A–1, the figure shows the distribution of the maximum sample size per site across all the RD models estimated in those studies. Each one of them reports several specifications, of which we consider the one with the largest number of observations per site. The histograms is composed of 17 observations because (i) Abdulkadiroglu et al. (2014) use entirely different samples for the Boston and New York City exam schools; (ii) for three studies we could not infer the number of sites from the information reported in the paper.

Full Sample, we keep a progressively smaller fraction of randomly selected observations in each site, separately by treatment status. When, at any round, a site would remain without treated or non-treated observations, we classify it as a "Panda" site to be "protected", and we keep it in the sample with at least 4 observations (at least two treated and two non-treated). At each round of the process, the theoretical fraction of observations that we keep is reduced by 5 percentage points until we reach a sample that would have 5% of the Full Sample observations if all sites were large enough to allow for such reduction. This procedure generates 19 samples (Sample95, Sample90, $\cdots$, Sample5), to which we add three samples corresponding respectively to theoretical fractions of 4, 3 and 2 percent of the Full Sample size. Descriptive statistics of some of these 22 reduced samples are reported in Table 1.

Thanks to this procedure, our analysis is relevant for researchers facing an average number of applicants per site that ranges between 716 (Sample95) and 15 (Sample2), keeping constant the number of sites at 1,729. As in Figure 1 shows, this is the relevant range for most of the papers that we reviewed. We cannot reduce the Full Sample to a theoretical size below 2% because our estimates would become too imprecise. However, the range of sample sizes that we consider is sufficient to infer (by extrapolation) what would happen when the number of applicants per site is even smaller than 15, as occasionally observed in the literature.

Table 1: Descriptive statistics of the Full Sample and of some representative reduced samples

	[1]	[2]	[3]	[4]	[5]	[6]
Theoretical fraction of Full Sample	2% Sample2	5% Sample5	10% Sample10	25% Sample25	50% Sample50	100% Full Sample
Total Sample size	25,686	62,418	124,257	310,110	620,074	1,238,418
Number of sites	1,729	1,729	1,729	1,729	1,729	1,729
Av. applicants per site	15	36	72	179	359	716
St.Dev of apps./site	10	25	51	128	256	511
% of “Panda” sites	42.5	24.8	15.6	7.0	3.4	0.0

Notes: Columns [1]-[5] report statistics for five reduced samples obtained from the Full Sample with the reduction procedure described in Section 8.2. The Full Sample has been constructed from the data of Pop-Eleches and Urquiola (2013) as described in Section 8.1. Column [6] reports the statistics for the Full Sample.

Three remarks about our sample-reduction procedure are in order. First, while keeping constant the number of sites, the procedure reduces the standard deviation of applicants per site (see the Online Appendix). This fact is important for the interpretation of our results, because this reduced heterogeneity contributes to decreasing the NP “bias”, thus favoring the NP estimator in the comparison with its alternatives.

Second, the identity of the marginal subject may change at any step of the process, because of re-sampling from the Full Sample. This happens when the subject located exactly at the cutoff in the Full Sample is dropped by chance in a smaller sample. We must therefore identify the new marginal subject (among the treated) in those sites where this is necessary at a given step of the process. However, once the new marginal subject is properly identified, this implication is immaterial.

Finally, the procedure is appropriate for a comparison of the different estimates at each given sample size, while the comparison of the estimates obtained for a given estimator at different sample sizes are not necessarily informative, particularly at small sample sizes. This is so because the presence of “Panda” sites implies that each sample corresponds to a different underlying population, featuring its own potentially different average treatment effect.

8.3 Estimates in the Full Sample

Following Gelman and Imbens (2019), we restrict our analysis to non-parametric versions of the estimators described in Sections 4–7. We obtain our estimates in Stata, using the `rdrobust` package developed by Calonico et al. (2017), after defining appropriately the running variable and the sample to implement the different estimation strategies. The comparison between these estimates in the Full Sample is displayed in columns [1]–[8] of Table 2. Columns [1] and [2] report the conventional NP estimates with and without marginal subjects located at the cutoff, respectively (see Section 4). Even at this very large sample size, the difference between these two estimates is quantitatively relevant: 0.014 when including marginal subjects vs 0.025 when excluding them. The optimal bandwidths computed by the Calonico et al. (2017) code are reported in the third row. The sample sizes in the last row differ because of the 1,729 marginal subjects dropped in column 2, which correspond to the number of sites reported in the next-to-last row.

Columns [3] and [4] report the corresponding SFE estimates, obtained using the same optimal bandwidths of their NP counterparts to make NP and SFE directly comparable. Independently of whether marginal subjects are included or not, as expected from our theoretical analysis in Section 5, these two SFE estimates are very similar (0.023 and 0.024, respectively). They are also very similar to the NP estimate without marginal subject, which confirms that when the sample size is very large, NP is "unbiased" for the causal parameter of interest only if marginal subjects are excluded. However, the efficiency gain of SFE suggests that this strategy dominates the NP estimator without marginal subjects even when the sample size is very large: the s.e. of the SFE estimates are 0.005 and 0.006 respectively in column [3] and [4], while for NP without marginal subjects the s.e. in column [2] is 0.008.

The first two rows in column [6] of Table 3 report the empirical counterparts of the two sides of sufficient condition (46) for the efficiency of SFE with respect to NP. The intraclass correlation of the composite error term $u_j + e_{ij}$ on the LHS of (46) is 0.47, while the intraclass correlation of the treatment indicator D_{ij} on the RHS is 0.006. The condition is therefore largely satisfied. For this reason, the SFE estimator is more precise than the NP estimator even at the very large sample size of the Full Sample. As explained in Section 5.2, a multi-cutoff RD setting in which condition (46) is satisfied is similar to the setting of a stratified RCT with a constant probability of treatment in each stratum and therefore the fixed effect transformation does not reduce in a relevant way the variability of the treatment dummy, while reducing the noise (Athey and Imbens, 2017).

Going back to Table 2, consider the estimates obtained with the SYM and SPLIT estimators described in Section 6. In column [5], the cutoff for the treated units (resp. non-treated) is set as the score of the marginally non-treated unit (resp. treated). In column [6], instead an artificial cutoff is set at half the distance between the score of the last treated and the score of the first non-treated observations. The resulting estimates are 0.023 and 0.027, respectively, but with standard errors that are in the ballpark of the NP ones and thus larger than those of the SFE estimates.

Finally, columns [7] and [8] of Table 2 report estimates based on defining the running variable in terms of ranks (RK). Also in this case, keeping marginal subjects in the analysis generates a "bias" comparable to the NP "bias" (0.013 versus 0.028 with and without marginal subjects, respectively). Moreover, in both cases the standard errors are large.

Recall that the Full Sample that we have constructed from the PEU original data, is not directly comparable to any of the samples used in the PEU study. However, for completeness, in column [9] of Table 2, we report the estimate that we obtain in the Full Sample with the parametric SFE estimator (without marginal subjects) used by PEU (first column of Panel B in Table 4 of their paper). Despite the sample differences described in Section 8.1, the two estimates are almost identical: 0.017 in the Full Sample and 0.018 in PEU, with a small standard error of 0.002 in both cases that is due to the parametric assumption. These results suggest that our Full Sample is not too different from the sample used by PEU for their corresponding estimates, despite our interventions on the original data.

Taking stock of the evidence described so far, two main messages emerge from Table 2 and confirm the theoretical predictions of Sections 4–7: first, at the very large sample size of PEU, Allocation rule 2 is not problematic in a multi-cutoff setting as long as marginal subjects located exactly at the cutoff are dropped from the analysis; second, the SFE estimator is more precise than any other estimator.

Table 2: Full sample estimates for the different estimators

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]	[9]
Estimator	NP	NP	SFE	SFE	SYM	SPLIT	RK	RK	PEU
Marginal subjects	Yes	No	Yes	No	Yes	Yes	Yes	No	No
Estimates	0.014	0.025	0.023	0.024	0.023	0.027	0.013	0.028	0.017
St. error	(0.007)	(0.008)	(0.005)	(0.006)	(0.009)	(0.008)	(0.009)	(0.008)	(0.002)
Optimal bandwith	0.228	0.177	0.228	0.177	0.136	0.161	6.968	8.456	n.a.
Number of sites	1,729	1,729	1,729	1,729	1,729	1,729	1,729	1,729	1,729
Sample size	1,238,418	1,236,689	1,238,418	1,236,689	1,238,418	1,238,418	1,238,418	1,236,689	1,236,689

Notes: The table reports non-parametric, Sharp RD estimates of the effect of being admitted to a higher-achievement high school in Romania on the grade of the Baccalaureate exam. Standard errors are clustered at the individual level, as in PEU. The data set is the Full Sample that we have constructed with the original data of Pop-Eleches and Urquiola (2013), as described in Section 8.1. Marginal subjects are the applicants located exactly at one of the 1,729 cutoffs in the Full Sample, because of Allocation rule 2. Therefore, the columns with and without marginal subjects differ by 1,729 observations. The RD estimates have been obtained with Local Linear Regressions using a uniform kernel and the optimal bandwidth from Calonico et al. (2014). The NP estimators are the sample analog of the Normalizing and Pooling estimands in equations (26) and (32), depending on whether marginal subjects are included in the sample or not. The SFE estimators are the sample analog of the Site Fixed Effect estimand in equation (35) with the weights for Allocation Rule (2) described in Section 5. The SYM and the SPLIT estimators are the sample analogs of the estimands described in Section 6. The RK estimators are the sample analog of the Rank Distance estimands described in equation (56). The last column reports the estimate that we obtain in the Full Sample with the parametric SFE estimator (without marginal subjects) used by PEU in the first column of Panel B in Table 4 of their paper.

Table 3: Sufficient condition for the efficiency of SFE with respect to NP, without marginal subjects, at different sample sizes

	[1]	[2]	[3]	[4]	[5]	[6]
Theoretical fraction	2%	5%	10%	25%	50%	100%
of Full Sample	Sample2	Sample5	Sample10	Sample25	Sample50	Full Sample
LHS of sufficient condition (46)	0.57	0.52	0.50	0.48	0.48	0.47
RHS of sufficient condition (46)	0.12	0.09	0.05	0.02	0.01	0.006
NP Estimates	0.009	−0.035	−0.019	0.002	0.018	0.025
NP Standard error	(0.053)	(0.036)	(0.022)	(0.012)	(0.008)	(0.008)
SFE Estimates	0.007	0.017	0.020	0.023	0.024	0.024
SFE Standard error	(0.034)	(0.021)	(0.016)	(0.008)	(0.005)	(0.006)

Notes: In the top panel, the table report the LHS and RHS of the sufficient condition (46) for the efficiency of SFE with respect to NP. Column [1]-[5] are for five reduced samples obtained from the Full Sample with the reduction procedure described in Section 8.2. Column 6 is for the Full Sample. The Full Sample has been constructed from the data of Pop-Eleches and Urquiola (2013) as described in Section 8.1. The bottom panel reports, for each selected sample, the corresponding NP and SFE estimates obtained without marginal subjects. For the Full Sample in column [6], these estimates are just replicated, for convenience of the reader, from columns [2] and [4] of Table 2, with their own standard errors clustered at the individual student level as in PEU. For the previous columns, we have first bootstrapped 100 reduced samples with replacement for each sample size. The estimates reported in the table are the averages of the 100 bootstrapped estimates for each sample size. The standard errors are the empirical standard deviations of the 100 bootstrapped estimates for each sample size. The full set of NP and SFE estimates without marginal subjects for all the reduced samples are plotted in the right panel of Figure 2, that will be described and discussed later.

In light of our theoretical analysis, neither of these conclusions may not hold at smaller sample sizes. As explained in Section 4, under Allocation rule 2 a necessary condition for the NP estimand to be “biased” is that there is heterogeneity across sites in the mean potential outcome of the treated at a cutoff, $\mathbb{E}[Y_{ij}(1)|0, j]$, as well as heterogeneity in the weight of each site in the left open neighborhood of a cutoff, $\frac{N_j f_j}{Nf}$.

The first two rows of Table 4 display the standard deviations of these quantities and confirm that such heterogeneity conditions hold in the Full Sample. In this case, the correlation between the two quantities indicates whether the size of the bias at smaller sample sizes might be quantitatively relevant. The third row shows that this correlation is indeed non-negligible, so we should expect a potentially important NP “bias” at smaller sample sizes, but not too small to eliminate the variability in the number of subjects per site.

As for the second conclusion, in principle sufficient condition (46) for the efficiency of SFE with respect to NP may not hold in the reduced samples, even if it holds in the Full Sample. Indeed, when the overall sample size is small our strategy to oversample small sites may induce a large reduction of the intra-class correlation of the error term. We next explore how the different estimators perform at the smaller sample sizes obtained by reducing the PEU sample in the way described in Section 8.2.

Table 4: Fundamentals of the bias components in the Full Sample

Parameter	Description	Estimate
$\text{SD}\left(\mathbb{E}[Y_{ij}(1)\|0, j]\right)$	Std. Dev. of mean potential outcome with treatment, at the cutoff	0.642
$\text{SD}\left(\frac{N_j f_j}{Nf}\right)$	Std. Dev. of relative site weight in $\mathbb{O}^-$	0.816
$\text{Corr}\left[\mathbb{E}[Y_{ij}(1)\|0, j], \frac{N_j f_j}{Nf}\right]$	Correlation between mean potential outcome with treatment, at cutoff, and the relative site weight in $\mathbb{O}^-$	0.490

Notes: The statistics reported in the rows of this table have been computed in the Full Sample described in the text, that we have constructed with the original data of Pop-Eleches and Urquiola (2013). The optimal bandwith for the calculation of these statistics is the one of the NP estimator including marginal subjects. The number of applicants is $N = 1,238,418$ and the number or sites is $J = 1,729$.

8.4 Estimates in the reduced samples

Figure 2 compares, at sample sizes that are smaller than the Full Sample, the NP and SFE estimation strategies. The horizontal axis measures the average number of subjects per site in the 22 reduced samples constructed with the procedure described in Section 8.2, which ranges between about 716 and 15. For each of these sample sizes, we bootstrapped 100 different samples, with replacement, and the vertical axis measures the average NP (triangle) and SFE (circle) estimates over these 100 bootstrap samples. The shaded areas in the figure describe the empirical 95% confidence intervals for each estimation strategy.[22] When the number of subjects per site becomes smaller than 50, the 95% confidence interval of the NP estimate becomes so large that its representation distorts the scale of the figure. For this reason, the shaded areas for these NP confidence intervals at the three smallest sample sizes are not represented in this figure (as well as in those that follow in this section), an issue that does not arise for the more precise SFE estimator.

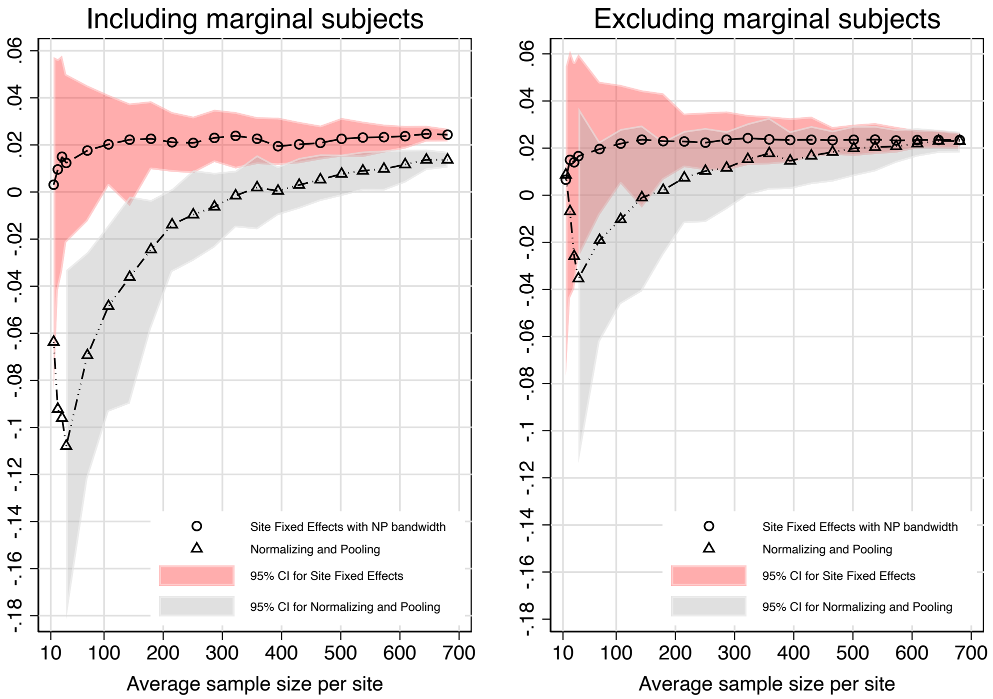
In the figure's left panel, the marginal subjects located exactly at each cutoff are included in the analysis. The SFE estimates (circles) are quite stable at a value of about 0.02 for all sample sizes larger than 100 subjects per site. This is about the same value obtained with the SFE estimator in the Full Sample (see column 3 of Table 2). In this range, the fraction of "Panda" sites is relatively low (see Table 1) and so the corresponding samples are roughly randomly selected from the Full Sample. It is therefore not surprising that the unbiased SFE estimator produces stable estimates in this range.

When the number of subjects per site falls below 50, the fraction of "Panda" sites increases substantially, becoming as high as 42.5% in the smallest sample (15 applicants per site; see Table 1). In this range of sample sizes, there is no reason to expect the treatment effect to be the same as the one estimated in the Full Sample, because these samples are not randomly smaller versions of the Full Sample. Indeed the SFE estimates decline to about 0.007 in the smallest sample.

The vertical difference between the SFE (circle) and the NP (triangle) estimates at each given sample size displayed in the left panel of Figure 2 is a measure of the "bias" of the

[22] The Online Appendix reports analogous figures in which the confidence intervals are obtained with the asymptotic approximation. The two approaches deliver very similar results.

Figure 2: Normalizing and Pooling (NP) vs Site Fixed Effects (SFE) estimates



Notes: This figure displays, for the reduced samples, the NP estimates (triangles) and the corresponding SFE estimates (circles) of the effect of being admitted to a higher-achievement high school in Romania on the grade of the Baccalaureate exam. All estimates are non-parametric, sharp RD. In the left panel marginal subjects (i.e., subjects located exactly at the cutoff), are included, while in the right panel they are dropped. The NP estimators are the sample analog of the Normalizing and Pooling estimand in equations (26) and (32), depending on whether marginal subjects are included or not. The SFE estimators are the sample analog of the Site Fixed Effect estimand in equation (35) with the weights for Allocation Rule 2 described in Section 5. The reduced samples have been obtained from the Full Sample constructed from the original data of Pop-Eleches and Urquiola (2013), with the procedure described in Sections 8.1 and 8.2. The RD estimates have been obtained with Local Linear Regressions using a triangular kernel and the optimal bandwidth from Calonico et al. (2014). The bandwith for the SFE estimator is the same as the optimal bandwith for the corresponding NP estimator. For each estimator, the shaded areas describe the 95% empirical confidence intervals of the estimates. To construct these confidence intervals we have first bootstrapped 100 samples with replacement for each reduced sample size. The boundaries of the 95% confidence intervals have then been set equal to the corresponding appropriate percentiles of the distribution of the 100 bootstrapped estimates.

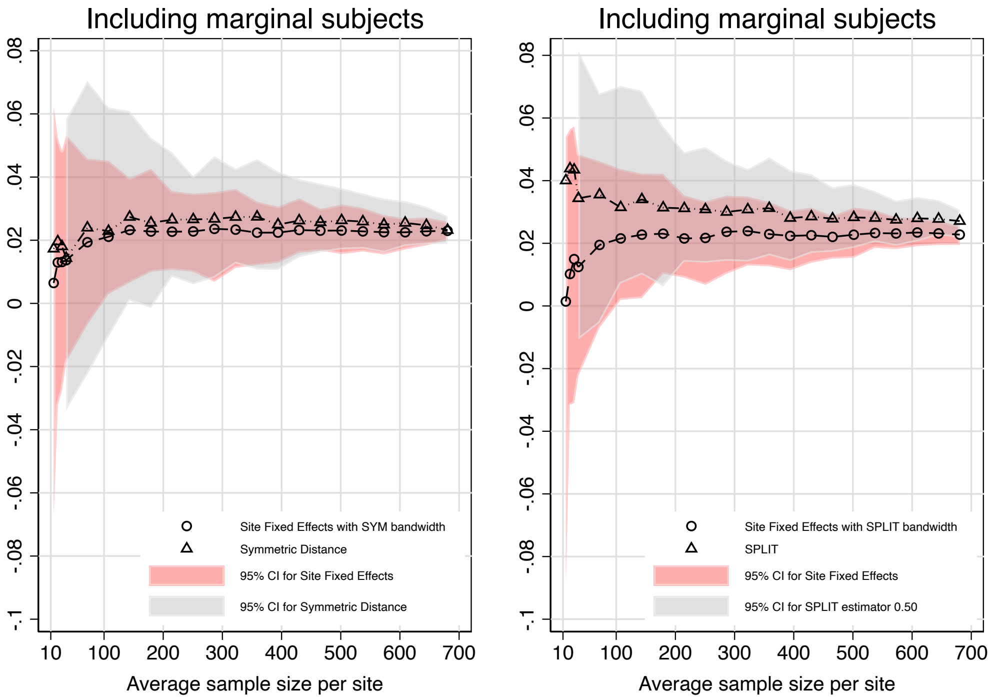
NP estimand when marginal subjects are included. Note that, as in Table 2, to make the NP and the SFE strategies directly comparable, the SFE estimates have been obtained with the optimal bandwidth of the NP estimate, as computed by the Calonico et al. (2017) code. This "bias", which is already sizeable in the Full Sample becomes even larger in absolute value at smaller sample sizes until the number of subjects per site is about 40. When the sample is further reduced, the absolute size of the "bias" declines (because of the reduced between-site variability in local size). However, at these small sample sizes the sampling variability of the NP estimator (not visible in the figure) increases dramatically.

To better understand the pattern of the "bias" (first increasing and then decreasing), recall that the reduction of the number of subjects per site *increases* the absolute size of the "bias" (conditional on the heterogeneity of sites with respect to their weight) but also reduces the heterogeneity of sites with respect to their weight because of the increasing fraction of "Panda" sites. This second effect, which in our application is quite important below 40 subjects per site, contributes to *reducing* the size of the "bias". However, researchers working in this range of small sample sizes – where site heterogeneity is necessarily limited – should not feel safe in using the NP estimator that includes marginal subjects, because this estimator is so imprecise that its estimates can take very different values.

The right panel of Figure 2 compares the NP and the SFE strategies when the marginal subjects located at each cutoff are excluded from the analysis. Dropping marginal subjects clearly helps reducing the absolute size of the "bias" at all sample sizes, but the qualitative pattern is the same. The NP point estimates turns from positive to negative when the number of subjects per site decreases below about 140. The "bias" is negative and grows in size until the number of subjects per site is reduced to about 50. Further sample reductions lead to a smaller "bias" because of the greater homogeneity in the numbers of subjects across sites, but the NP estimator remains very imprecise also at very small sample sizes.

Thus, even if dropping marginal subjects helps reducing the NP "bias", it does *not* eliminate it at intermediate values of applicants per site (see Section 4.2); moreover, the SFE estimates are considerably more precise (see Section 5.2). To better assess this precision gain, Table 3 reports for a subset of reduced samples, the values of the LHS and RHS of sufficient condition (46) in these samples. The table also reports the corresponding NP and SFE estimates obtained without marginal subjects. The standard errors are the empirical standard deviations of the 100 bootstrapped estimates for each sample size. The LHS of (46) grows from 0.47 to 0.57 when sample size declines from 100% to 2% of the Full Sample. These values are considerably larger than the corresponding ones for the RHS, which approaches its theoretical maximum of 0.125 (see Section 5.2) in the case of the smallest sample size. Here the average number of subjects included in the optimal bandwidth is about 4, so the condition is always satisfied and the SFE standard errors are always smaller than the NP ones. The gap in precision increases with the reduction of the sample size.

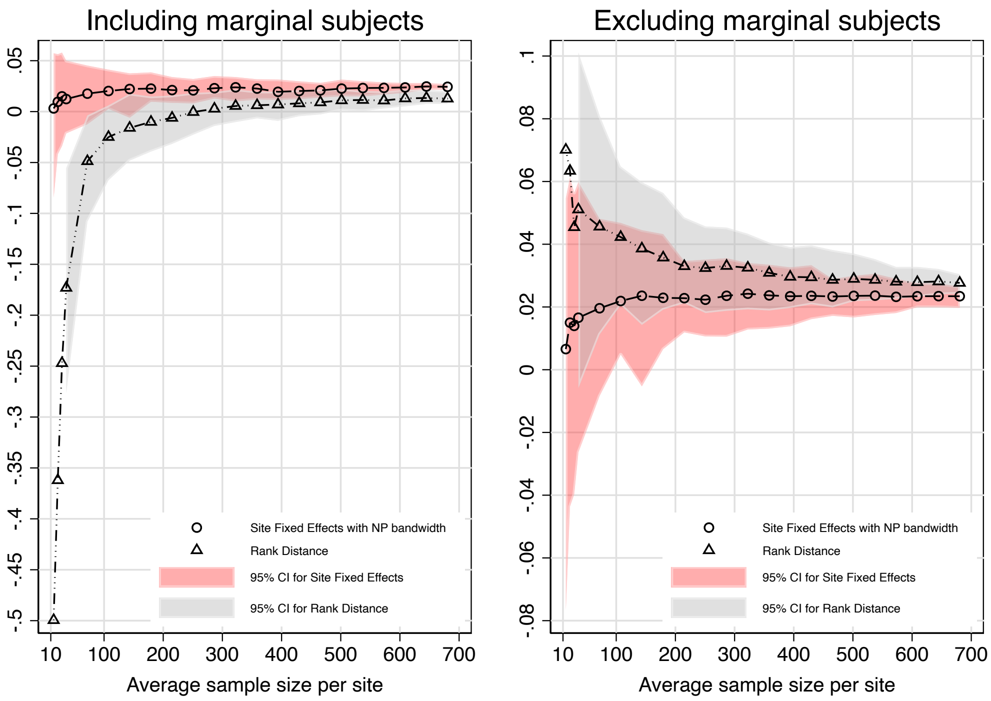
Figure 3: Estimates based on redefining the cutoff vs Site Fixed Effects (SFE) estimates



Notes: The left panel of this figure displays the SYM estimates (triangles) and the corresponding SFE estimates (circles) in the reduced samples. The right panel displays instead the SPLIT estimates (triangles) and the corresponding SFE estimates (circles) in the reduced samples. All estimates are for the effect of being admitted to a higher-achievement high school in Romania on the grade of the Baccalaureate exam, and are non-parametric, sharp RD. The SYM and the SPLIT estimators are the sample analogs of the estimands described in Section 6. The SFE estimators are the sample analog of the Site Fixed Effect estimand in equation (35) with the weights for Allocation Rule 2 described in Section 5. In both panels marginal subjects (i.e., subjects located exactly at the cutoff), are included. The reduced samples have been obtained from the Full Sample constructed with the original data of Pop-Eleches and Urquiola (2013), with the procedure described in Sections 8.1 and 8.2. The RD estimates have been obtained with Local Linear Regressions using a triangular kernel and the optimal bandwidth from Calonico et al. (2014). The bandwith for the SFE estimator is the same as the optimal bandwith for the corresponding SYM or SPLIT estimator. For each estimator, the shaded areas describe the 95% empirical confidence intervals of the estimates. To construct these confidence intervals we have first bootstrapped 100 samples with replacement for each reduced sample size. The boundaries of the 95% confidence intervals have then been set equal to the corresponding appropriate percentiles of the distribution of the 100 bootstrapped estimates.

The SYM and SPLIT estimators, based on re-defining the cutoff (see Section 6), are compared with their SFE counterparts in Figure 3. Also in these cases, to make the two solutions directly comparable, the SFE estimates have been obtained with the same optimal bandwidth of the corresponding competing estimators. Interestingly, for both estimators the "bias" with respect to SFE is positive, in contrast with the negative NP "bias". The left panel of this figure describes the performance of the SYM estimator based on redefining the cutoff that is relevant for the treated (resp. non-treated) as the score of the marginally non-treated (resp. treated). This estimator performs remarkably well in terms of "bias" but is less precise than the corresponding SFE, particularly at small sample sizes. When instead the cutoff is set at half the distance between the scores of the last treated and the first non-treated observations (SPLIT), the estimator performs less well in general and particularly when the number of subjects per site is smaller than about 100. Again, at all sample sizes this estimator is considerably less precise than its SFE counterpart.

Figure 4: Rank distance (RK) and corresponding Site Fixed Effects (SFE) estimates

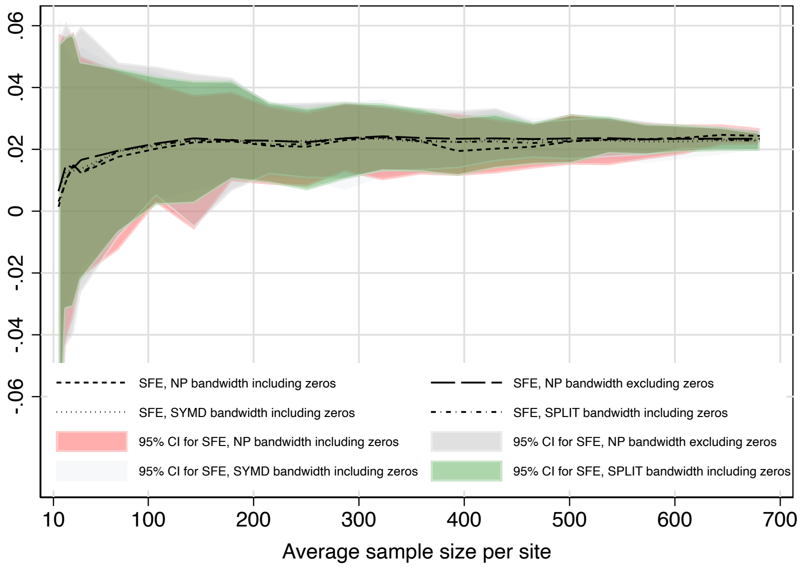


Notes: This figure displays the RK estimates (triangles) and the SFE estimates (circles) of the corresponding NP estimator (see Figure 2) in the reduced samples. All estimates are for the effect of being admitted to a higher-achievement high school in Romania on the grade of the Baccalaureate exam. In the left panel marginal subjects (i.e., subjects located exactly at the cutoff), are included, while in the right panel they are dropped. All estimates are non-parametric, sharp RD. The RK estimators are the sample analog of the Rank Distance estimands described in equation (56). The SFE estimators are the sample analog of the Site Fixed Effect estimand in equation (35) with the weights for Allocation Rule 2 described in Section 5. The reduced samples have been obtained from the Full Sample constructed with the original data of Pop-Eleches and Urquiola (2013), with the procedure described in Sections 8.1 and 8.2. The RD estimates have been obtained with Local Linear Regressions using a triangular kernel and the optimal bandwidth from Calonico et al. (2014). The bandwith for the SFE estimator is the same as the optimal bandwith for the corresponding NP estimators. For each estimator, the shaded areas describe the 95% empirical confidence intervals of the estimates. To construct these confidence intervals we have first bootstrapped 100 samples with replacement for each reduced sample size. The boundaries of the 95% confidence intervals have then been set equal to the corresponding appropriate percentiles of the distribution of the 100 bootstrapped estimates.

Figure 4 considers the estimators based on defining the running variable in terms of ranks (RK). Since the metric of the running variable is now different, for SFE we use the optimal bandwidth of the NP estimator. That is, we compare the RK estimators (with and without marginal subjects) with the SFE estimators that we have used as counterparts of the NP estimators. Note that the range of values on the vertical axis is considerably larger than in previous figures and so although the graphic display may suggest that the bias of these estimators is small (at least at large sample sizes) in fact it is always sizable. At the small sample sizes that are more typical in the literature (i.e., below 200 subjects per site) the bias is huge.

Finally, Figure 5 compares all the four SFE estimators that have been displayed in the previous figures. Note that they differ only in the value of the bandwidth. This comparison is instructive because it shows that, at any sample size, the SFE estimator generates very similar estimates, which are therefore robust to the choice of alternative bandwidths. It also shows that these solutions are similarly very precise.

Figure 5: Comparison of the Site Fixed Effects (SFE) estimates

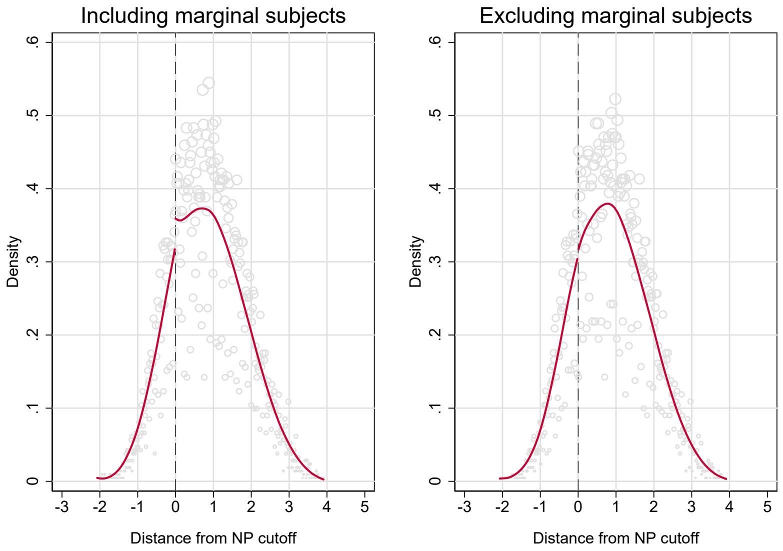


Notes: This figure displays the four non-parametric SFE estimates in the reduced samples corresponding to the two NP, the SYM and the SPLIT estimators displayed in previous figures 2–4. All estimates are for the effect of being admitted to a higher-achievement high school in Romania on the grade of the Baccalaureate exam. Marginal subjects (i.e., subjects located exactly at the cutoff), are dropped only for the NP estimator that excludes them. All the SFE estimators are the sample analogs of the Site Fixed Effect estimand in equation (35) with the weights for Allocation Rule 2 described in Section 5. They differ because of the optimal bandwiths that are set equal to those of the corresponding NP, SYM or SPLIT estimators. The reduced samples have been obtained from the Full Sample constructed with the original data of Pop-Eleches and Urquiola (2013), with the procedure described in Sections 8.1 and 8.2. The RD estimates have been obtained with Local Linear Regressions using a uniform kernel and the optimal bandwidth from Calonico et al. (2014). For each estimator, the shaded areas describe the 95% empirical confidence intervals of the estimates. To construct these confidence intervals we have first bootstrapped 100 samples with replacement for each reduced sample size. The boundaries of the 95% confidence intervals have then been set equal to the corresponding appropriate percentiles of the distribution of the 100 bootstrapped estimates.

9 Conclusions: drawing lessons for the practitioner

The first implication of Allocation rule 2 is that at the outset of a multi-cutoff research project a practitioner would detect a point discontinuity in the empirical density of the running variable at the NP cutoff, which is induced by the marginal subjects located exactly at the cutoff. Following PEU's analysis, we report in Figure 6 this density, estimated using the survey sample for the 2005-2007 admission cohorts. The right panel excludes applicants who are at zero distance from the cutoff, and coincides with Figure A.7 in the Online Appendix of PEU. In this case, as the authors report, the McCrary (2008) test does not detect a significant discontinuity (the log difference in height is 0.074, s.e. 0.058). The left panel shows instead that if marginal subjects are included in the sample, then the density exhibits a visible and significant discontinuity (here the log difference in height is 0.230, s.e. 0.054).

Figure 6: Density in Pop-Eleches and Urquiola's (2013) survey data



Notes: The figure shows the density of the running variable after normalizing and pooling all cutoffs, using the survey sample for the 2005-2007 admission cohorts in Pop-Eleches and Urquiola (2013) – a total of 11,931 observations. The density in the left panel uses all observations. The density in the right panel excludes 93 observations located exactly at the cutoff, and coincides (up to bandwidth) with Figure A.7 in the Online Appendix to Pop-Eleches and Urquiola (2013). The size of circles is proportional to the number of observations in each bin, and the fit is obtained from Local Linear Regressions on each side of the cutoff, using a uniform kernel and optimal bandwidth from Calonico et al. (2014).

It is possible that the observation of such a "jump" discourages the practitioner from continuing the project. However, this discontinuity should *not* be perceived as a problem as long as potential outcomes are continuous at each cutoff, which is the case in PEU.[23] When facing this situation, the researcher should adopt one of the empirical strategies that we have shown to be effective in bypassing the problem, most notably the SFE strategy. At the same time, the researcher should be aware that simply dropping the marginal subjects located at the cutoff may not be a solution because, even if it restores the continuity of the density as shown in the right panel of Figure 6, it does not always solve the identification problem described in Section 4.2. This failure is for example evident for the NP estimator in the right panel of Figure 2.

In order to illustrate these results more clearly, our analysis has introduced several simplifications in the original PEU data. In these final remarks we clarify that these simplifications do not preclude the generality of our results and we discuss what practitioners should do in relation to these issues. For example, we have dropped from the PEU sample all the sites that did not display rationing, while, typically, the practitioner will have data in which some sites may feature no rationing (see Section 8.1). Moreover, even in the absence of sites without rationing, it may happen that some sites do not have subjects on one side of the cutoffs *within the bandwidth* (and specifically below the cutoff under Allocation rule 2). The NP estimator uses all sites to compute the above and below average outcomes, including sites in which no proper counterfactual comparison for the treated units is possible. SFE instead forces the analyst to use only sites with observations locally on both sides, i.e. the only sites in which the identification of the treatment effect is feasible.

We have also eliminated ties from the PEU data set by adding a tiny noise to the running variable (see again Section 8.1). Also in this case the practitioner will instead most likely face multi-cutoff settings with Allocation Rule 2 and ties in the running variable. It is easy to see, from the definition of the estimators in Sections 4 and 5, that ties away from the cutoff pose no problem to any of the estimators that we have considered. Ties are instead potentially more problematic for the comparison of estimators with and without marginal subjects if they are located exactly at the cutoff. Again, in order not to distort the comparison between these estimators, we have eliminated ties but in practice their presence does not change in any important way our analysis. The SFE estimator remains the safest option in the

[23] See Hahn et al. (2001) and Choi and Lee (2019).

presence of ties at the cutoff.

Finally, in the comparison between each estimator and its SFE counterpart, we have forced the SFE bandwidth to be the same as the focal estimator. But what is the optimal bandwidth that a practitioner should use to adopt the SFE strategy independently of other options? To our knowledge, there is no theory to select an optimal bandwidth for the SFE estimator in the context that we study. Hence, the safe empirical strategy is to check the robustness of SFE to alternative choices of the bandwidth, as we do in Figure 5.

Summing up, we have compared in this paper the performance of different estimators in RD designs with multiple cutoffs in which a marginally exposed subject is located exactly at each cutoff. This occurs whenever a fixed number of treatment slots is allocated starting from the subject with the highest (or lowest) value of the score, until exhaustion. Our theoretical analysis, illustrated with the data of Pop-Eleches and Urquiola (2013), suggests that a fixed effect estimation strategy is by far the safest option and it is also likely to be more precise.

Appendix

Table A–1: Multi-cutoff RD studies that feature one subject located at each cutoff

Paper	Field	Sites	N sites
A. Referenced by Cattaneo et al. (2016):			
Boas and Hidalgo (2011)	Pol. Sci.	Election/Coalition	40,341
Boas, Hidalgo and Richardson (2014)	Pol. Sci.	Election/Coalition	*Many*
Goodman (2008)	Education	School district/Year	867
Hainmueller and Kern (2008)	Pol. Sci.	Election/District	*Many*
Kane (2003)	Education	Grant Rank/Year	4
Kendall and Rekkas (2012)	Pol. Sci.	Election/District	10,889
Klasnja (2015)	Pol. Sci.	Election/City	$\approx$ 9,000
Trounstine (2011)	Pol. Sci.	Elections/City	*Many*
Uppal (2009)	Pol. Sci.	Elections	24,592
B. Additional references:			
Abdulkadiroglu, Angrist and Pathak (2014)	Education	School/Year	12-30
Bedoya, Gonzaga, Herrera and Espinoza (2019)	Education	School	482
Black, Galdo and Smith (2007)	Labor	Empl. office/Week	1,107
Cohodes and Goodman (2014)	Education	School district/Year	1,156
David, Smith-McLallen and Ukert (2019)	Health	Outreach wave	10
Estrada and Gignoux (2017)	Education	School	634
Kirkeboen, Leuven and Mogstad (2016)	Education	University track/Year	3,360
Francis-Tan and Tannuri-Pianto (2018)	Education	University track	318
Fort, Ichino and Zanella (2020)	Education	Daycare program	546
McEachin, Domina and Penner (2020)	Education	School/Year	753
Pop-Eleches and Urquiola (2013)	Education	School/Year	1,984
Wu, Wei, Zhang and Zhou (2019)	Education	School/Year	8

Notes: The table lists recent papers in the Regression Discontinuity (RD) literature featuring multiple thresholds and where the design implies that each threshold is the value of the running variable for a marginal subject exposed to the treatment, so that there is at least one observation located exactly at each cutoff. Panel A lists papers already reviewed in Cattaneo et al. (2016). Panel B is our update of their review. Following Cattaneo et al. (2016) we use *Many* to refer "to examples based on vote shares, where the cutoff is a continuous random variable; in these cases, the number of cutoffs is related to the number of effective parties".

References

Abdulkadiroglu, A., Angrist, J.D., Pathak, P.A., 2014. The Elite Illusion: Achievement Effects at Boston and New York Exam Schools. Econometrica 82, 137–196.

Angrist, J.D., Pischke, J.S., 2008. Mostly Harmless Econometrics: An Empiricist's Companion. Princeton University Press.

Athey, S., Imbens, G., 2017. Chapter 3 - the econometrics of randomized experimentsa, in: Banerjee, A.V., Duflo, E. (Eds.), Handbook of Field Experiments. North-Holland. volume 1 of *Handbook of Economic Field Experiments*, pp. 73 – 140. URL: http://www.sciencedirect.com/science/article/pii/S2214658X16300174, doi:https://doi.org/10.1016/bs.hefe.2016.10.003.

Barreca, A.I., Lindo, J.M., Waddell, G.R., 2015. Heaping-Induced Bias in Regression-Discontinuity Designs. Economic Inquiry .

Bedoya, M., Gonzaga, B., Herrera, A., Espinoza, K., 2019. Setting an example? spillover effects of Peruvian magnet schools .

Bertanha, M., 2020. Regression Discontinuity Design with Many Thresholds. Journal of Econometrics - forthcoming.

Black, D.A., Galdo, J., Smith, J.A., 2007. Evaluating the worker profiling and reemployment services system using a regression discontinuity approach. The American Economic Review 97, 104 –107.

Boas, T.C., Hidalgo, F.D., 2011. Controlling the airwaves: Incumbency advantage and community radio in brazil. American Journal of Political Science 55, 869–885.

Boas, T.C., Hidalgo, F.D., Richardson, N.P., 2014. The spoils of victory: Campaign donations and government contracts in brazil. Journal of Politics 76, 415–429.

Calonico, S., Cattaneo, M.D., Farrell, M., Titiunik, R., 2017. rdrobust: Software for Regression Discontinuity Designs. The Stata Journal 17, 372–404.

Calonico, S., Cattaneo, M.D., Titiunik, R., 2014. Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs. Econometrica 82, 2295–2326.

Cattaneo, M.D., Titiunik, R., Vazquez-Bare, G., Keele, L., 2016. Interpreting regression discontinuity designs with multiple cutoffs. The Journal of Politics 78.

Choi, J., Lee, M., 2019. Continuity of the Running Variable Density Is Neither Necessary Nor Sufficient for Regression Discontinuity Validity. mimeo.

Cohodes, S.R., Goodman, J.S., 2014. Merit aid, college quality, and college completion: Massachusetts' Adams Scholarship as an in-kind subsidy. American Economic Journal: Applied Economics 6, 251–285.

David, G., Smith-McLallen, A., Ukert, B., 2019. The effect of predictive analytics-driven interventions on healthcare utilization. Journal of Health Economics 64, 68 – 79.

de Chaisemartin, C., Behaghel, L., 2020. Estimating the effect of treatments allocated by randomized waiting lists. Econometrica - forthcoming.

Estrada, R., Gignoux, J., 2017. Benefits to elite schools and the expected returns to education: Evidence from Mexico City. European Economic Review 95, 168–194.

Fort, M., Ichino, A., Zanella, G., 2020. Cognitive and noncognitive costs of day care at age 0–2 for children in advantaged families. Journal of Political Economy 128, 158–205.

Francis-Tan, A., Tannuri-Pianto, M., 2018. Black Movement: Using discontinuities in admission to study the effects of college quality and affirmative action. Journal of Development Economics 135, 97–116.

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

Goodman, J., 2008. Who merits financial aid? Massachusetts' Adams Scholarship. Journal of Public Economics 92, 2121 – 2131.

Hahn, J., Todd, P., Van der Klaauw, W., 2001. Identification and Estimation of Treatment Effects with a Regression Discontinuity Design. Econometrica 69, 201–209. Notes and Comments.

Hainmueller, J., Kern, H.L., 2008. Incumbency as a source of spillover effects in mixed electoral systems: Evidence from a regression-discontinuity design. Electoral Studies 27, 213–227.

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

Kendall, C., Rekkas, M., 2012. Incumbency advantages in the canadian parliament. Canadian Journal of Economics 45, 1560–1585.

Kirkeboen, L.J., Leuven, E., Mogstad, M., 2016. Field of Study, Earnings, and Self-Selection. The Quarterly Journal of Economics 131, 1057–1111.

Klasnja, M., 2015. Corruption and the incumbency disadvantage: Theory and evidence. Journal of Politics 77, 928–942.

McCrary, J., 2008. Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test. Journal of Econometrics 142, 698–714.

McEachin, A., Domina, T., Penner, A., 2020. Heterogeneous effects of early algebra across California middle schools. Journal of Policy Analysis and Management 0, 1–20. Forthcoming.

Pearl, J., 2014. Comment: Understanding Simpson's Paradox. The American Statistician 68, 8–13.

Pop-Eleches, Urquiola, M., 2013. Going to a Better School: Effects and Behavioural Responses. American Economic Review 103, 1289–1324.

Trounstine, J., 2011. Evidence of a local incumbency advantage. Legislative Studies Quarterly 36, 255–280.

Uppal, 2009. The disadvantaged incumbents: estimating incumbency effects in indian state legislatures. Public Choice 138, 9–27.

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