Electronic Journal of Statistics
Vol. 17 (2023) 243–290
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2102

# Kernel regression analysis of tie-breaker designs*

**Dan M. Kluger and Art B. Owen**

*Department of Statistics, Stanford University, Stanford CA, 94305*
*e-mail:* kluger@stanford.edu; owen@stanford.edu

**Abstract:** Tie-breaker experimental designs are hybrids of Randomized Controlled Trials (RCTs) and Regression Discontinuity Designs (RDDs) in which subjects with moderate scores are placed in an RCT while subjects with extreme scores are deterministically assigned to the treatment or control group. In settings where it is unfair or uneconomical to deny the treatment to the more deserving recipients, the tie-breaker design (TBD) trades off the practical advantages of the RDD with the statistical advantages of the RCT. The practical costs of the randomization in TBDs can be hard to quantify in generality, while the statistical benefits conferred by randomization in TBDs have only been studied under linear and quadratic models. In this paper, we discuss and quantify the statistical benefits of TBDs without using parametric modelling assumptions. If the goal is estimation of the average treatment effect or the treatment effect at more than one score value, the statistical benefits of using a TBD over an RDD are apparent. If the goal is nonparametric estimation of the mean treatment effect at merely one score value, we prove that about 2.8 times more subjects are needed for an RDD in order to achieve the same asymptotic mean squared error. We further demonstrate using both theoretical results and simulations from the Angrist and Lavy (1999) classroom size dataset, that larger experimental radii choices for the TBD lead to greater statistical efficiency.

**MSC2020 subject classifications:** Primary 62K99; secondary 62G20, 62G08.
**Keywords and phrases:** Causal inference, experimental design, hybrid experiments, local linear regression, regression discontinuity designs.

Received March 2022.

## Contents

1 Introduction . . . . . 244
2 Literature review . . . . . 246
2.1 Regression discontinuity methods . . . . . 246
2.2 Experimental design . . . . . 248
2.3 Tie-breakers . . . . . 248
3 Causal estimand and problem formulation . . . . . 250
4 Asymptotic mean square optimal error . . . . . 253
5 Variance comparisons at fixed $h > \Delta$ . . . . . 258

*This work was supported by the U.S. National Science Foundation under grants IIS-1837931 and DMS-2152780 and by Stanford University's SGF and SIGF fellowships.

6 Classroom size data . . . . . 263
7 Discussion . . . . . 268
A Proof of Lemma 2 . . . . . 269
B Proof of Theorem 1 . . . . . 276
C Extensions to assignment probability $p \neq 1/2$ . . . . . 277
D Bandwidth that removes leading order bias . . . . . 280
E Proof of Proposition 2 . . . . . 284
F Proof of Proposition 3 . . . . . 285
Acknowledgments . . . . . 288
References . . . . . 288

## 1. Introduction

In this paper we study a nonparametric regression approach to tie-breaker studies. In the settings of tie-breaker studies, there is a costly treatment while the control is inexpensive or even free. In addition, an investigator can decide how to allocate the costly treatment using a priority ordering on the subjects. The priority ordering could be based on how deserving of the treatment each subject is, or based on how strongly each subject is expected to respond to the treatment. Examples include offering scholarships or school placement to students (Angrist, Autor and Pallais, 2020), offering a drug rehabilitation program to people of varying needs (Cappelleri and Trochim, 2003), or assigning interventions to reduce risk factors for child abuse and neglect (Krantz, 2022). In these settings a randomized controlled trial (RCT) is inappropriate because it is extremely inefficient economically, or even ethically questionable. The natural, even automatic, approach to settings like this is to rank the subjects $i = 1, \dots, N$ according to their value of a running variable $x_i$ and assign the treatment to only those subjects with the highest values of $x_i$. For simplicity one can assume that the number of subjects to treat is fixed and that the treatment is offered to subject $i$ if and only if $x_i > t$ for some threshold $t$.

The problem with a deterministic treatment based on $x_i$ is that it complicates causal inference of the effect of the treatment. One can use regression discontinuity analysis (Thistlethwaite and Campbell, 1960, Cattaneo and Titiunik, 2022) but the regression discontinuity design (RDD) is known to give treatment effect estimates a very high variance. See, for example, Gelman and Imbens (2019). In a parametric regression model the treatment and running variable are highly correlated, making for an inefficient design (Goldberger, 1972, Jacob et al., 2012). In a tie-breaker design (TBD), the subjects are given the treatment with a probability that increases with $x_i$. The top ranked subjects get the treatment with probability one, the bottom ranked subjects do not get the treatment and subjects in between are randomly assigned to either treatment or control. The tie-breaker design interpolates between two extremes: the RCT and the RDD, trading off statistical efficiency with the short term economic value of aligning treatment to the running variable.

We believe that there are many more good uses for TBDs. Many companies interact electronically with their customers and partners. Perks, such as service upgrades, can easily be assigned with some randomization. Because the treatment is costly, it is important to evaluate the treatment efficacy later. This provides a strong motivation to introduce randomization. When the perk is simply a gift to some subjects there is less ethical concern over whether it goes to the most loyal customers or introduces some randomization. We also expect that tie-breakers will be useful in evaluating governmental programs such as the one in Krantz (2022), as well as educational programs such as those in Angrist, Autor and Pallais (2020).

TBDs have been primarily studied as experimental designs using parametric regression modeling assumptions. While the design literature focuses on parametric models, the RDD literature primarily uses nonparametric regression methods. In this paper, we quantify the statistical gains to be obtained by conducting a TBD instead of an RDD using nonparametric regression.

Our main theoretical contributions are as follows. We study a kernel weighted local linear regression with a slope and intercept for both treated and control subjects and bandwidth $h$. The RDD can consistently estimate the treatment effect only at $x = t$, and so we focus our comparison at that point. We find an expression for the optimal bandwidth for estimating the treatment effect at $x = t$ under the TBD. We then compare the optimal mean squared error at $x = t$ for the two designs. For the popular triangular kernel, a TBD reduces the asymptotic mean squared error (AMSE) by a factor of about 2.27 compared to an RDD of the same sample size, $N$. For other popular kernels, the AMSE is reduced by a slightly greater factor. In this setting, the AMSE decreases proportionally to $N^{-4/5}$, and using $N$ points in an RDD is comparable to using only $0.36N$ points in a TBD. The asymptotic analysis has a bandwidth that converges to zero. Since this convergence is at the very slow $N^{-1/5}$ rate, we cannot assume that in practice $h$ will be small enough such that subjects without randomized treatment are discarded. Therefore, we also compare the designs when $h$ is fixed and large enough to include nonnrandomized subjects in the regression. In this setting, the efficiency ratio, which we define as the relative variance of treatment effect estimators under the two designs, can be as large as four. For a fixed bandwidth and for the triangular and boxcar kernels, we also find that the efficiency ratio is monotone non-decreasing in the proportion of subjects who are given a randomized treatment assignment.

A further advantage of the TBD is that it can give consistent nonparametric estimates of the treatment effect for any value of $x$ in the randomization window. It can also be used to estimate the average causal effect over that window. These additional advantages are described more explicitly in Section 3.

An outline of this paper is as follows. Section 2 reviews the literature on tie-breakers as well as the much larger literature on RDDs. In Section 3, we define a causal parameter of interest that can be used to compare the TBD to the RDD. We also introduce the causal identification assumptions needed and the local linear regressions used for estimating that parameter. In Section 4, we compare the mean squared error (MSE) in asymptotically optimal estimation

of our causal parameter of interest under an RDD to that under a TBD. In that asymptotic setting, the optimal bandwidth decreases at the slow $O(N^{-1/5})$ rate and then the local linear regression is eventually supported entirely in the experimental region of the TBD. In Section 5, we investigate another regime where the bandwidth $h$ is fixed and is assumed to be larger than the radius $\Delta$ of the experimental region. For this setting, deferring an investigation of the bias to Appendix D, we study the variance of our estimator as a function of $\Delta$ and find the efficiency ratio to be monotone in $\Delta$ for the triangular and boxcar kernels. Section 6 shows how one can compute efficiency ratios empirically using one's actual assignment variable levels, focusing on the Israeli classroom size data from Angrist and Lavy (1999) as an example. The curves of the empirical efficiency ratios are quite similar to the ones obtained theoretically. Section 7 presents a discussion. Appendix C extends our results to TBDs in which each subject in the experimental group is given the treatment with probability $p \neq 1/2$. Appendices A, B, E, and F contain some of our proofs.

## 2. Literature review

Here we survey the small TBD literature and some recent developments in the much larger RDD literature. We also note connections to the experimental design literature. Most of the TBD literature has focused on global parametric models. Those are also the dominant model for experimental design. The TBD is usually compared to the RDD, for which nonparametric models are the norm.

### *2.1. Regression discontinuity methods*

Here we present some concepts from the regression discontinuity literature drawing heavily on Cattaneo and Titiunik (2022). We begin with a setting where there is a running variable $x_i$ (also called a score or index) for subject $i = 1, \dots, N$. Subjects with $x_i > t$ are given the treatment and others get the control. The treatment levels are typically $T_i \in \{0, 1\}$ with $T_i = 0$ being the control. For the TBD setting it is more convenient to use $Z_i \in \{-1, 1\}$ with $Z_i = -1$ indicating the control. The potential outcomes for subject $i$ are $Y_{i+}$ if treated and $Y_{i-}$ for control.

There are two main approaches to RDD in causal inference, continuity-based and local randomization-based. The continuity-based approach assumes that the mean response for treated subjects is continuous in $x$ as is that for control subjects. If the mean response for all subjects shows a discontinuity at $x = t$, then the magnitude of this discontinuity is defined to be the causal effect of treatment on subjects at $x = t$, and one then considers how to estimate that effect. The version from Hahn, Todd and der Klaauw (2001) has IID tuples $(x_i, Y_{i+}, Y_{i-})$ where $Y_i$ equals $Y_{i+}$ for $x_i > t$ and $Y_i$ equals $Y_{i-}$ otherwise. The treatment effect is then

$$\tau = \lim_{x \downarrow t} \mu_+(x) - \lim_{x \uparrow t} \mu_-(x)$$

where $\mu_\pm = \mathbb{E}(Y_\pm \mid X = x)$. We will work primarily with a superpopulation setting where the subjects in the study are sampled from a joint distribution. Cattaneo and Titiunik (2022) discuss this setting along with some other settings that focus on causal inference for the given subjects.

The local randomization approach from Cattaneo, Frandsen and Titiunik (2015) assumes the existence of a window $\mathcal{W} = [t-h, t+h]$ such that for $x \in \mathcal{W}$ the treatment variable is 'as good as randomized'. In the local randomization approach we assume **a)** that the joint distribution of the $Z_i$ for $x_i \in \mathcal{W}$ is known, and, **b)** that the potential outcomes $(Y_{i+}, Y_{i-})$ are independent of $x_i$. In particular, both mean responses must be constant functions of $x \in \mathcal{W}$. A variant of local randomization has the treatment based on a threshold of $x$ where $x$ is a noisy version of a latent variable $u$ (Eckles et al., 2020) and where we have outside information under which the probability of treatment given $u$ is known.

Both frameworks have challenges. The obvious difficulty with local randomization is choosing the window $\mathcal{W}$ (or knowing the treatment probability given $u$). A smaller window provides a smaller dataset to use while making the window larger will normally increase the discrepancy between the model and the ground truth. The TBD can be viewed as a strategy to impose by design the first assumption in the local randomization approach while not making the second assumption.

The challenge in the continuity framework is in estimating the necessary limits. In a parametric model, those limits are estimated from all the data but have a bias due to lack of fit of the parametric model. As a result, nonparametric regression methods based on local polynomial models are favored. The challenge there is that one must choose a bandwidth $h$, analogously to the window size from the local randomization framework. Because the mean responses are only locally polynomial we must contend with a bias-variance tradeoff in estimating the limits.

Our theoretical and numerical results compare the TBD to the RDD in the continuity framework. We think that this is the more likely alternative analysis for our motivating applications if randomization had not been used, because the local randomization assumptions do not seem natural in those applications. We focus on the accuracy of point estimation. There is also a large literature on constructing confidence intervals around the RDD estimate (see Cattaneo and Titiunik (2022)). We describe some of those concepts in this paper, but we do not develop confidence intervals for the TBD due to space constraints.

There are many different settings where treatments depend in a discontinuous way on $x$. In a sharp design, the treatment is $Z_i = 1$ if and only if $x_i > t$. In a fuzzy design, the assignment to treatment or control might not perfectly match $x_i > t$ versus $x_i \leqslant t$, for reasons beyond the control of the investigator. For instance there may be subjects that do not comply with their assigned treatment. A related issue is that some subjects might be able to manipulate their value of the running variable in order to get (or avoid) the treatment. Rosenman and Rajkumar (2019) discuss some ways to counter that problem. In the settings we consider, the investigator has control of the treatment and so we study the sharp design. We also do not address issues of subject compliance, as

we suppose that the effect of the treatment assignment or the intent to treat is of sufficient interest to the investigator.

The RDD setting has been extended well beyond the simple framework described above. There are versions with treatments at more than two levels as well as versions with continuous treatments. The cutoff can be defined in terms of a vector of covariates yielding a discontinuity set of dimension one less than the vector has. The treatment discontinuity could be defined by geographical boundaries. There are multi-cutoff settings where subject $i$ gets the treatment when $x_i > t_i$. There are models where it is the derivative of $\mathbb{E}(Y \,|\, x, Z = 1) - \mathbb{E}(Y \,|\, x, Z = -1)$ that has a step discontinuity at $t$. For discussion and references to the variants above, see Cattaneo and Titiunik (2022).

### 2.2. Experimental design

While we study TBDs in comparison to regression discontinuity, they can also be considered within an experimental design framework such as the covariate-dependent designs considered by Metelkina and Pronzato (2017). That paper emphasizes sequential problems, and like most of the design literature it works primarily with parametric models. They find conditions where the treatment policies converge to optimal deterministic functions of a covariate vector. The TBD does not use deterministic allocations which is an advantage if the response distribution is subject to change between experiments.

Experimental design, especially in a sequential setting is closely related to bandit methods. We are motivated by problems where the responses $Y_i$ arrive too slowly for bandit methods to be suitable. In a business setting, the responses may arrive after a year or calendar quarter while the effect of a scholarship on graduation rates can only be seen years later.

### 2.3. Tie-breakers

The simplest tie-breaker design replaces the threshold $t$ by two thresholds $t \pm \Delta$. Subjects with $x_i > t + \Delta$ get the treatment, subjects with $x_i < t - \Delta$ get the control and other subjects are randomized to either treatment or control. The simplest choice has

$$\Pr(Z_i = 1 \,|\, x_i) = \begin{cases} 0, & x_i < t - \Delta, \\ \frac{1}{2}, & |x_i - t| \leqslant \Delta \\ 1, & x_i > t + \Delta. \end{cases} \tag{1}$$

Campbell (1969) describes the $\Delta = 0$ version of this design. Some subjects are exactly at the threshold $t$ and then randomization breaks the ties among them. Boruch (1975) considers positive values of $\Delta$ such that differences in the running variable among subjects with $|x - t| \leqslant \Delta$ are essentially arbitrary because $x$ is an imperfect measure. Abdulkadiroğlu et al. (2022) study the New York school

system that breaks ties among applicants by lottery, or standardized test, or audition, depending on the program. We only consider randomized tie-breaking.

Goldberger (1972) considers a simple two line regression model that in our notation is

$$Y_i = \beta_1 + \beta_2 x_i + \beta_3 Z_i + \beta_4 x_i Z_i + \varepsilon_i \tag{2}$$

for IID errors $\varepsilon_i$ with mean zero and variance $\sigma^2$. He finds that an RDD estimates these coefficients with a variance that is asymptotically $\pi/(\pi-2) \approx 2.75$ times as large as it would be under an RCT. His setting has Gaussian $x_i$ and $\beta_4 = 0$. Jacob et al. (2012) generalizes the above model to polynomials of degree two or three in $x$ with or without interactions between $x$ and $Z$. Their Table 6 shows that an RCT is 4 times as efficient as the RDD for a uniformly distributed running variable with $t$ at the midpoint of its range. They also provide similar efficiency estimates for other polynomial models for both uniform and Gaussian $x$ and include settings where $t$ is not at the median of the distribution of $x$.

The above comparisons of RCTs to RDDs do not include tie-breakers. Cappelleri, Darlington and Trochim (1994) compare small, medium and large randomization windows in which 20%, 35% and 50%, respectively, of the subjects get a randomized treatment. They tabulate the sample sizes needed to attain a certain level of statistical power for three treatment effect sizes in these TBDs as well as in an RDD and in an RCT. All designs had half of the subjects getting the treatment. The running variable $x$ was normally distributed. The model was (2) with $\beta_4 = 0$, making the treatment effect constant. The power calculations were done by Monte Carlo sampling. The required sample sizes became smaller with increased randomization at any level of power and effect size.

Owen and Varian (2020) work out the asymptotic variance of $\hat{\beta}$ in the model (2) as a function of $\Delta$. They consider both $\mathbb{U}[-1,1]$ and $\mathcal{N}(0,1)$ distributions for $x$ and a threshold $t$ at the median of $x$'s distribution. The estimated treatment effect is $2(\hat{\beta}_3+\hat{\beta}_4 x)$ and they find for uniform $x$ that this estimate has asymptotic variance proportional to $16(1+3x^2)/[1+3\Delta^2(2-\Delta^2)]$ where $\Delta = 0$ describes the RDD and $\Delta = 1$ is the RCT. This decreases monotonically in $\Delta$ while increasing monotonically in $|x|$.

They also consider the opportunity cost of experimentation compared to the RDD. For $x_i \sim \mathbb{U}[-1,1]$ the expected value of $\sum_{i=1}^N Y_i$ is approximately $(\beta_1 + \beta_4(1-\Delta^2)/2)N$. If larger $Y_i$ are better and $\beta_4 > 0$ then the opportunity cost grows proportionally to $\beta_4\Delta^2$. They discuss how one might trade off this opportunity cost against statistical efficiency.

The TBD has so far been analyzed for simpler methods than the RDD has. This can be understood by comparing their workflows. In a TBD we measure $x_i$, then sample $Z_i$ and then some time later observe $Y_i$. For an RDD we usually get $(x_i, Z_i, Y_i)$ all at once. The investigator planning a TBD only has $x_i$, must decide how to assign the $Z_i$, and may not know what model will be fit later, and then chooses some specific model to design for. When one studies the TBD theoretically, one does not even have the $x_i$ and then it is natural to assume a distribution for them. The TBD is prospective while the RDD is retrospective.

When vectors $\boldsymbol{x}_i$ of covariates are available, Owen and Varian (2020, Section 8) describe how to investigate numerically the efficiency of a TBD that fits a regression model on some collection of features of $\boldsymbol{x}_i$ that interact with $Z_i$ where the treatment window is based on a linear combination of $\boldsymbol{x}_i$.

Morrison and Owen (2022) study multiple regression for a tie-breaker in a regression model $Y_i = \boldsymbol{x}_i^\mathsf{T}\beta + Z_i\boldsymbol{x}_i^\mathsf{T}\gamma + \varepsilon_i$ with $\Pr(Z_i = 1) = p_i \in [0, 1]$. They study a prospective $D$-optimality criterion that maximizes the determinant of $\mathbb{E}(\mathcal{X}^\mathsf{T}\mathcal{X})$, where $\mathcal{X}$ is the (random) design matrix built from $\boldsymbol{x}_i$ and $Z_i$. For any known $\boldsymbol{x}_i$, the finite sample optimal $p_i$ can be computed by convex optimization. For as yet unobserved $\boldsymbol{x}_i$ the prospective $D$-optimality criterion averages over both random $Z_i$ and random $\boldsymbol{x}_i$ from an assumed distribution for $\boldsymbol{x}_i$. For random $\boldsymbol{x}_i$, they study a three level tie-breaker with running variable $\boldsymbol{x}_i^\mathsf{T}\eta$ and treatment probabilities 0, 0.5 and 1.

Owen and Varian (2020) consider replacing the simple trichotomy (1) by various sliding scales where $\Pr(Z_i = 1 \,|\, x_i)$ is a monotone function of $x_i$. They find no advantage to such alternatives when $x_i$ has a symmetric distribution about $t$ and half the subjects are treated. Li and Owen (2022) revisit that problem for the two line model and find optimal designs for general $x_i$ distributions and general fractions of treated subjects without assuming that half of the subjects will be treated. These optimal designs can greatly improve upon the design defined in (1). They still have $\Pr(Z_i = 1 \,|\, x_i)$ as piecewise constant functions of $x_i$. If we impose monotonicity $\Pr(Z_i = 1 \,|\, x_i) \geqslant \Pr(Z_{i'} = 1 \,|\, x_{i'})$ whenever $x_i \geqslant x_{i'}$, then only two treatment probability levels are needed.

A limitation of previous comparisons between RDDs and TBDs is that they all assume parametric regression models for the response $Y_i$. We compare them using local linear regression. For simplicity, we restrict our attention to the three level version of the TBD in (1).

Our model assumes an additive error on top of smooth functions of $x$ for the data points where $|x - t| > \Delta$. When $h > \Delta$, the causal estimate we consider merges deterministic and randomized treatment allocations and then cannot be analyzed in a potential outcomes framework. It is common in causal inference to ignore such data. For instance, a rule of thumb in Crump et al. (2009) is to omit data where the treatment probability is outside $[0.1, 0.9]$. Asymptotically, $h < \Delta$ and then a potential outcomes analysis is available. Otherwise, to stay within the potential outcomes framework, one must choose between ignoring some data and using the additive error model like we do.

## 3. Causal estimand and problem formulation

Throughout the text we will compare the TBD to the RDD. In our comparison, we will define $t$ to be the putative RDD threshold and $\Delta$ to be the experimental radius, and we consider allocation of the treatments to the $N$ subjects according to the 3-level tie-breaker design (1).

Next we discuss the estimands of interest. For each subject we consider the assignment variable $X \in \mathbb{R}$, the treatment $Z \in \{-1, 1\}$ and two potential out-

comes: $Y_+ = Y(Z = 1)$ and $Y_- = Y(Z = -1)$. Defining

$$\mu_+(x) \equiv \mathbb{E}\big(Y_+ \,|\, X = x\big) \quad \text{and} \quad \mu_-(x) \equiv \mathbb{E}\big(Y_- \,|\, X = x\big), \tag{3}$$

the treatment effect at $X = x$ is

$$\tau(x) = \mathbb{E}(Y_+ - Y_- \,|\, X = x) = \mu_+(x) - \mu_-(x).$$

If the investigator chooses an RDD with a threshold at $t$, then under certain regularity conditions, the causal estimand

$$\tau_{\text{thresh}} \equiv \tau(t)$$

can be consistently estimated. In particular, we assume IID samples (or sufficiently weak dependence between the samples) and that:

(i) The density $f(\cdot)$ of the assignment variable $X$ is continuous at $t$ with $f(t) > 0$.
(ii) The conditional mean functions $\mu_+$ and $\mu_-$ in (3) have at least 3 continuous derivatives in an open neighborhood of $t$.
(iii) The conditional variance functions $\sigma^2_\pm(x) \equiv \text{Var}\big(Y_\pm \mid X = x\big)$ are both bounded in a neighborhood of $t$ and continuous at $t$.

Under these conditions, $\tau_{\text{thresh}}$ can be consistently estimated by local linear regression with $O_p(N^{-2/5})$ errors (Imbens and Kalyanaraman, 2012). On the other hand, the conditions above do not suffice to let an RDD consistently estimate $\tau(x)$ for any $x \neq t$.

If an investigator runs a TBD with $\Delta > 0$, then assumptions like those above replacing $t$ by $x$ allow consistent estimation of $\tau(x)$ for any $x \in (t - \Delta, t + \Delta)$. Furthermore, as long as $\text{var}(Y_\pm) < \infty$,

$$\tau_{\text{ATE}}(\Delta) \equiv \mathbb{E}(\tau(X) \,|\, t - \Delta < X < t + \Delta)$$

can be consistently estimated with error $O_p(N^{-1/2})$ in a TBD without requiring assumptions (i), (ii), and (iii).

The discussion above leaves open the possibility that an RDD could be better than a TBD when estimating $\tau_{\text{thresh}}$. Therefore, for the remainder of the paper our primary focus will be on showing that even if the only goal is estimating $\tau_{\text{thresh}}$, it is still beneficial to run a TBD rather than an RDD. Our other focus will be to show that when the only goal is to estimate $\tau_{\text{thresh}}$, it is beneficial to pick a larger $\Delta$ in the experimental design stage when the option is available. Picking a larger $\Delta$ has other benefits as well such as making $\tau(x)$ identifiable for more values of the assignment variable and making $\tau_{\text{ATE}}(\Delta)$ more representative of the overall population and easier to estimate. Naturally, there are non-statistical reasons to keep $\Delta$ smaller.

### *Local linear estimation*

In keeping with current RDD practice, we suppose that under an RDD $\tau_{\text{thresh}}$ will be estimated with local linear regression. In particular, we assume that a

parameter vector $\beta$ defined by

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^4}{\arg\min} \sum_{i=1}^{N} K\Big(\frac{x_i - t}{h}\Big)\big(Y_i - (\beta_1 + \beta_2 x_i + \beta_3 Z_i + \beta_4 x_i Z_i)\big)^2 \tag{4}$$

will be fit for some symmetric kernel function $K(\cdot) \geqslant 0$ and bandwidth parameter $h > 0$, and that $\tau_{\text{thresh}}$ will be estimated with

$$\hat{\tau}_{\text{thresh}} = 2\hat{\beta}_3 + 2\hat{\beta}_4 t. \tag{5}$$

While this formulation of estimating $\hat{\tau}_{\text{thresh}}$ may be less familiar than the approach of fitting separate local linear regressions for the treatment and control groups, it is easy to check that the two formulations yield the same estimator.

Throughout the paper, we suppose that under a TBD, $\hat{\tau}_{\text{thresh}}$ will also be estimated using local linear regression according to (4) and (5). We do not use the same bandwidth $h$ for the TBD and RDD. Indeed, in the next section we see that the optimal bandwidth choice (in terms of AMSE) is different for the two designs.

Because kernels with unbounded support are not typically used in RDD analysis (Cattaneo and Titiunik (2022)), we only consider kernels with bounded support. We assume without loss of generality, that the kernel is supported on $[-1, 1]$. We have a special interest in a uniform (boxcar) kernel $K_{\text{BC}}(x) = 1_{|x| \leqslant 1}$ because it is a popular kernel choice and is a local version of the regression model (2). We are also interested in a triangular spike kernel $K_{\text{TS}}(x) = (1-|x|)_+$ where $z_+ = \max(0, z)$. This kernel was shown by Cheng, Fan and Marron (1997) to optimize a bias-variance tradeoff for extrapolation from $x_i > t$ to $\mathbb{E}(Y \,|\, x = t)$ and has been advocated for RDD analysis by Imbens and Kalyanaraman (2012) and Calonico, Cattaneo and Titiunik (2014) among others.

The local linear regression estimator from (5) has a bias and variance that both depend on the bandwidth $h$. Larger $h$ typically bring greater bias because the true regression is not precisely linear over a region centered on $t$. Smaller $h$ bring greater variance because then fewer data points are in the regression. Imbens and Kalyanaraman (2012) develop a method for choosing the bandwidth $h$ that is asymptotically mean squared optimal for the RDD. In the next section, we compare the AMSE of the TBD with that of the RDD, when each of them has their asymptotically optimal bandwidth choice.

In this paper, we focus on the accuracy of the estimated treatment effect. The RDD literature includes several papers devoted to the construction of confidence intervals. There it is necessary to account for the bias in a local polynomial regression. A simple approach is to choose $h$ to undersmooth the regression function, resulting in a bias of lower order than the standard error, and this simplifies confidence interval construction. Undersmoothing, however, brings less accuracy (Calonico, Cattaneo and Titiunik, 2014). See Calonico, Cattaneo and Farrell (2019) for a discussion of bandwidth choices to optimize estimation, or optimize confidence interval construction, or to get robust (asymptotically valid) confidence intervals using the bandwidth that is optimal for estimation.

## 4. Asymptotic mean square optimal error

In this section, we demonstrate the advantage of the TBD over the RDD when each design's bandwidth is chosen to minimize the AMSE in the estimation of $\tau_{\text{thresh}} = \mu_+(t) - \mu_-(t)$. Following Imbens and Kalyanaraman (2012), we assume the following more general regularity conditions for estimating the causal effect at $X = t$:

(i) The triples $(X_i, Y_{i+}, Y_{i-})$ for $i = 1, \ldots, N$ are IID.
(ii) The distribution of $X_i$ has density $f(\cdot)$, which is continuously differentiable at $t$ with $f(t) > 0$.
(iii) Conditional means $\mu_\pm(\cdot)$ both have at least three continuous derivatives in an open neighborhood of $t$, with the $k$'th derivatives at $t$ denoted $\mu_\pm^{(k)}(t)$.
(iv) The kernel $K(\cdot)$ is nonnegative, symmetric, bounded, has support $[-1, 1]$, is continuous on its support, and is strictly positive somewhere.
(v) The conditional variances $\sigma_\pm^2(x) \equiv \text{var}(Y_{i\pm} \,|\, X_i = x)$ are both bounded in an open neighborhood of $t$ and are continuous and strictly positive at $t$.
(vi) $\mu_+^{(2)}(t) \neq \mu_-^{(2)}(t)$.

Under an RDD, $Z_i = 1$ if $X_i > t$ and is $-1$ otherwise, so these assumptions imply Assumptions 3.1–3.6 that Imbens and Kalyanaraman (2012) make for an RDD. To allow for analysis in the TBD setting, our assumptions (i)–(vi) are slightly stronger than those in Imbens and Kalyanaraman (2012). For example, unlike in our Assumption (iii), Imbens and Kalyanaraman (2012) make no assumptions on $\mu_+(\cdot)$ in the interval $(-\infty, t)$ or on $\mu_-(\cdot)$ in the interval $(t, \infty)$. Regarding assumption (vi), Imbens and Kalyanaraman (2012) also consider the case where $\mu_+^{(2)}(t) = \mu_-^{(2)}(t)$ and show that in this case, their proposed method of estimating $\tau_{\text{thresh}}$ has error $O_p(N^{-3/7})$ rather than $O_p(N^{-2/5})$. We do not consider the case where $\mu_+^{(2)}(t) = \mu_-^{(2)}(t)$ in detail for the TBD as the result should be similar to that for the RDD and is of less interest for our head-to-head comparison of TBD with RDD.

Because our assumptions (i)–(vi) imply Assumptions 3.1–3.6 in Imbens and Kalyanaraman (2012) for an RDD, if we let

$$\tilde{\nu}_j \equiv \int_0^\infty u^j K(u)\,\mathrm{d}u \quad \text{and} \quad \tilde{\pi}_j \equiv \int_0^\infty u^j K^2(u)\,\mathrm{d}u$$

for $j \in \mathbb{N}$, and let

$$\tilde{C}_1 \equiv \frac{1}{4}\Big(\frac{\tilde{\nu}_2^2 - \tilde{\nu}_1\tilde{\nu}_3}{\tilde{\nu}_0\tilde{\nu}_2 - \tilde{\nu}_1^2}\Big)^2 \quad \text{and} \quad \tilde{C}_2 \equiv \frac{\tilde{\nu}_2^2\tilde{\pi}_0 - 2\tilde{\nu}_1\tilde{\nu}_2\tilde{\pi}_1 + \tilde{\nu}_1^2\tilde{\pi}_2}{(\tilde{\nu}_2\tilde{\nu}_0 - \tilde{\nu}_1^2)^2}, \tag{6}$$

and define

$$\text{AMSE}_{\text{RDD}}(h, N) \equiv \tilde{C}_1\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2 h^4 + \frac{\tilde{C}_2}{Nh}\Big(\frac{\sigma_+^2(t) + \sigma_-^2(t)}{f(t)}\Big), \tag{7}$$

then Lemma 3.1 of Imbens and Kalyanaraman (2012) holds. We reproduce the statement of this lemma below.

**Lemma 1.** *Under Assumptions (i)–(vi), if an* RDD *determines the treatment assignment and both* $h \to 0$ *and* $Nh \to \infty$ *as the number of samples* $N \to \infty$*, then the mean squared error in estimating* $\tau_{\text{thresh}}$ *is given by*

$$\text{MSE}_{\text{RDD}}(h, N) = \text{AMSE}_{\text{RDD}}(h, N) + o_p\Big(h^4 + \frac{1}{Nh}\Big), \tag{8}$$

*and the asymptotically optimal bandwidth, defined by* $\arg\min_h \text{AMSE}_{\text{RDD}}(h, N)$ *is given by*

$$h_{\text{opt,RDD}}(N) = \Big(\frac{\tilde{C}_2}{4\tilde{C}_1}\Big)^{1/5}\Big(\frac{\sigma_+^2(t) + \sigma_-^2(t)}{f(t)\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2}\Big)^{1/5} N^{-1/5}. \tag{9}$$

*Proof.* Imbens and Kalyanaraman (2012, Lemma 3.1). □

Because we wish to compare the RDD to the tie-breaker design, we derive a similar result for the asymptotic MSE for the tie-breaker design. The TBD counterparts to the RDD quantities above are

$$\nu_j \equiv \int_{-\infty}^{\infty} u^j K(u)\,\mathrm{d}u \quad \text{and} \quad \pi_j \equiv \int_{-\infty}^{\infty} u^j K^2(u)\,\mathrm{d}u \tag{10}$$

for $j \in \mathbb{N}$,

$$C_1 \equiv \frac{1}{4}\Big(\frac{\nu_2^2 - \nu_1\nu_3}{\nu_0\nu_2 - \nu_1^2}\Big)^2 \quad \text{and} \quad C_2 \equiv \frac{2\nu_2^2\pi_0 - 4\nu_1\nu_2\pi_1 + 2\nu_1^2\pi_2}{(\nu_2\nu_0 - \nu_1^2)^2}, \tag{11}$$

and

$$\text{AMSE}_{\text{TBD}}(h, N) \equiv C_1\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2 h^4 + \frac{C_2}{Nh}\Big(\frac{\sigma_+^2(t) + \sigma_-^2(t)}{f(t)}\Big). \tag{12}$$

**Lemma 2.** *Under Assumptions (i)–(vi), if a* TBD *with a fixed experimental radius* $\Delta > 0$ *determines the treatment assignment and both* $h \to 0$ *and* $Nh \to \infty$ *as the number of samples* $N \to \infty$*, then the mean squared error in estimating* $\tau_{\text{thresh}}$ *is given by*

$$\text{MSE}_{\text{TBD}}(h, N) = \text{AMSE}_{\text{TBD}}(h, N) + o_p\Big(h^4 + \frac{1}{Nh}\Big), \tag{13}$$

*and the asymptotically optimal bandwidth, defined by* $\arg\min_h \text{AMSE}_{\text{TBD}}(h, N)$ *is*

$$h_{\text{opt,TBD}}(N) = \Big(\frac{C_2}{4C_1}\Big)^{1/5}\Big(\frac{\sigma_+^2(t) + \sigma_-^2(t)}{f(t)\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2}\Big)^{1/5} N^{-1/5}. \tag{14}$$

*Proof.* See Appendix A. □

The proof of this lemma is very similar to the proof of Lemma 3.1 in Imbens and Kalyanaraman (2012), from their appendix. Instead of pointing to their proof and noting the parts of their proof that differ in the tie-breaker design

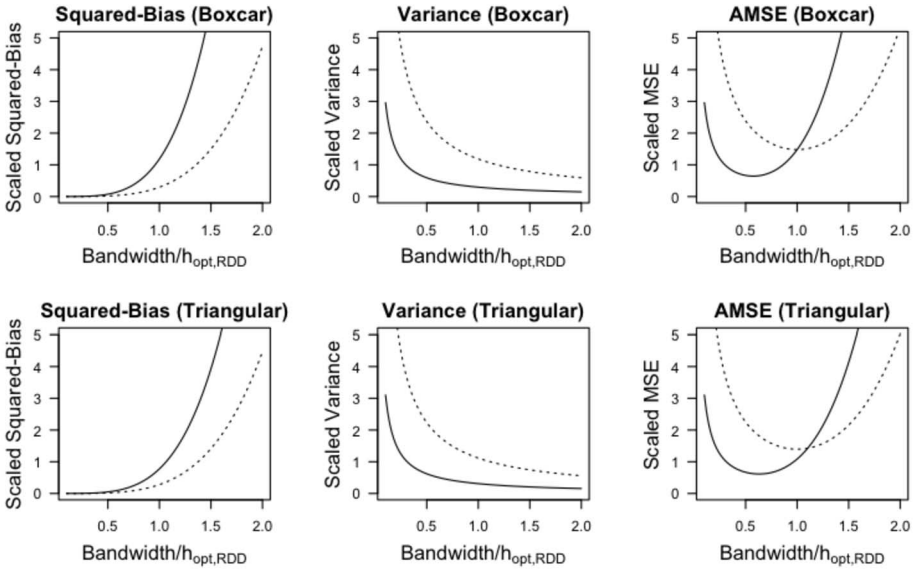



**Fig 1:** A comparison of the asymptotic bias-variance tradeoff for the regression discontinuity design (dotted lines) versus for the tie-breaker design (solid lines). The x-axes are in units of asymptotically MSE optimal bandwidth for RDD given at (9) while the y-axes are the leading order terms in units of $\alpha N^{-4/5}$ where $N$ is the sample size and $\alpha$ is a constant given in (15) that depends on properties of the joint distribution of $(X, Y, Z)$ in a neighborhood of the cutoff.

setting, we write out the proof of Lemma 2 in Appendix A to ensure there are no subtle issues with using their proof in the tie-breaker design setting.

The leading order MSE formulas are derived by evaluating and summing the leading order terms for both the squared-bias and the variance. In formulas (7) and (12) for the leading order MSE, the first term gives the leading order squared-bias while the second term gives the leading order variance. See formulas (36) and (37) for explicit calculations of the leading order bias and variance in the TBD case, and see the formulas for 'B' and 'V' in the appendix of Imbens and Kalyanaraman (2012) for explicit calculations of these quantities in the RDD case. It is not surprising that the formulas for the leading order squared-bias, variance and MSE are different for the two design types because for an RDD, estimation of $\tau_{\text{thresh}}$ involves estimation of the mean functions at a boundary point, whereas for a TBD, estimation of $\tau_{\text{thresh}}$ involves estimation of mean functions at an interior point.

In Figure 1, we plug in scalar multiples of the optimal bandwidth for the RDD given in Lemma 1 to the first and second terms of formulas (7) and (12) to visualize the trade-off for the leading order squared-bias and variance in a tie-breaker design compared to a regression discontinuity design. The formulas

simplify when defining the quantity

$$\alpha \equiv \frac{5}{4}\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^{2/5}\Big(\frac{\sigma_+^2(t) + \sigma_-^2(t)}{f(t)}\Big)^{4/5}, \tag{15}$$

which does not depend on $h$, $N$ or the kernel choice.

In practice, the optimal bandwidth is not known and must be estimated. For both the RDD and the TBD, the optimal bandwidth depends on the quantity

$$\gamma \equiv \Big(\frac{\sigma_+^2(t) + \sigma_-^2(t)}{f(t)\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2}\Big)^{1/5} \tag{16}$$

which must be estimated from the observed data. We consider the regularized estimator for $\gamma$ of Imbens and Kalyanaraman (2012). We take the estimated optimal bandwidth $\hat{h}_{\text{opt}}$ proposed in their Section 4.2 and set $\hat{\gamma}_{\text{RDD}} = (4\tilde{C}_1/\tilde{C}_2)^{1/5}\hat{h}_{\text{opt}}N^{1/5}$. It can be seen from the proof of Theorem 4.1 in Imbens and Kalyanaraman (2012) that under assumptions (i)–(vi), $\hat{\gamma}_{\text{RDD}} \xrightarrow{p} \gamma$. In the TBD case, we know a consistent estimator of $\gamma$ exists. For example, if we let $\hat{\gamma}_{\text{TBD,naive}}$ be an estimator of $\gamma$ that is constructed similarly to $\hat{\gamma}_{\text{RDD}}$ using only the subset of the data which looks like an RDD, $\hat{\gamma}_{\text{TBD,naive}} \xrightarrow{p} \gamma$. Of course such an estimator of $\gamma$ is inefficient; in practice one should instead use an estimator of $\gamma$ that does not throw out all of the control samples for which $x > t$ and all of the treated samples for which $x < t$. For our theoretical comparison of TBDs with RDDs, we are not concerned with the actual form of $\hat{\gamma}_{\text{TBD}}$ as long as it is consistent. Therefore, in the TBD case we will let $\hat{\gamma}_{\text{TBD}}$ be any estimator that satisfies $\hat{\gamma}_{\text{TBD}} \xrightarrow{p} \gamma$. We make a few remarks about estimation of $\gamma$ in the TBD setting in the discussion section.

To compare the AMSE for the RDD versus the TBD, we will assume that if the investigator were to run an RDD and were seeking mean squared optimal estimation of $\tau_{\text{thresh}}$, they would ultimately use the bandwidth

$$\hat{h}_{\text{opt,RDD}}(N) = \Big(\frac{\tilde{C}_2}{4\tilde{C}_1}\Big)^{1/5}\hat{\gamma}_{\text{RDD}}N^{-1/5}, \tag{17}$$

where $\hat{\gamma}_{\text{RDD}}$ is the consistent estimator for $\gamma$ described above and $\tilde{C}_j$ are defined at (6). We will also assume that if the investigator were to run a TBD seeking mean squared optimal estimation of $\tau_{\text{thresh}}$, they would ultimately use the bandwidth

$$\hat{h}_{\text{opt,TBD}}(N) = \Big(\frac{C_2}{4C_1}\Big)^{1/5}\hat{\gamma}_{\text{TBD}}N^{-1/5}, \tag{18}$$

where $\hat{\gamma}_{\text{TBD}}$ is any consistent estimator of $\gamma$ and $C_j$ are defined at (11).

The following theorem compares the RDD with $N$ points to a TBD with $\theta N$ points for some $\theta > 0$. We will use the value of $\theta$ that provides equal MSEs for estimation of $\tau_{\text{thresh}}$ as a metric to compare the two designs.

**Theorem 1.** *Let $\theta > 0$ be a constant. Under assumptions (i)–(vi), as $N \to \infty$*

$$\frac{\mathrm{MSE}_{\mathrm{RDD}}\big(\hat{h}_{\mathrm{opt,RDD}}(N), N\big)}{\mathrm{MSE}_{\mathrm{TBD}}\big(\hat{h}_{\mathrm{opt,TBD}}(\theta N), \theta N\big)} \xrightarrow{p} \theta^{4/5}\Big(\frac{\tilde{C}_1\tilde{C}_2^4}{C_1C_2^4}\Big)^{1/5} \tag{19}$$

*holds for any tie-breaker design of the form (1) with $\Delta > 0$.*

*Proof.* See Appendix B. □

Theorem 1 uses the assumption that $\Pr(Z_i = 1 \mid x_i) = 1/2$ for $x_i$ in the randomization window. If $\sigma_+^2(t) \neq \sigma_-^2(t)$, then we might prefer to offer the treatment with probability $p \neq 1/2$. In Appendix C, we study a treatment probability $p \in (0, 1)$. When $p = \sigma_+(t)/(\sigma_+(t) + \sigma_-(t))$, the asymptotic MSE is minimized, though an investigator would also want to account for the cost of the treatment. If one chooses $p$ using poor prior estimates of $\sigma_\pm$ it is possible that the resulting TBD will have a higher asymptotic MSE than the RDD. However, for any of the kernels in Table 1, one can protect against that by choosing $p \in [0.18, 0.82]$.

### *Asymptotic MSE comparison for some specific kernels*

We now use Theorem 1 to compare the MSE in estimating $\tau_{\mathrm{thresh}}$ for the RDD versus the TBD, under optimal bandwidth choices for various kernels of interest. See Table 1. If an investigator is deliberating between an RDD with $N$ samples versus conducting a TBD (for a fixed $\Delta > 0$) with $N$ samples, and either experimental design is to be analyzed with the asymptotically optimal bandwidth choice for the prespecified kernel, then the ratio of the MSEs will converge in probability to $\big((\tilde{C}_1\tilde{C}_2^4)/(C_1C_2^4)\big)^{1/5}$ as $N \to \infty$. Using formulas (6) and (11), the fourth column of Table 1 gives the value of the quantity $\big((\tilde{C}_1\tilde{C}_2^4)/(C_1C_2^4)\big)^{1/5}$ rounded to 2 decimal places. For the boxcar and triangular kernels respectively, this quantity is precisely $64^{1/5}$ and $60.46618^{1/5}$ without rounding.

It is also interesting to consider the quantity given by

$$\theta_* = \frac{C_1^{1/4}C_2}{\tilde{C}_1^{1/4}\tilde{C}_2}. \tag{20}$$

As a result of Theorem 1, an experimental designer deciding to use a TBD rather than an RDD would only need to collect $\theta_*$ times as many samples in order to achieve the same asymptotic MSE in estimating $\tau_{\mathrm{thresh}}$.

Table 1 shows that the kernel choice has a remarkably small impact on the relative benefit of using a TBD rather than an RDD to estimate $\tau_{\mathrm{thresh}}$. It is well known in the usual kernel smoothing setting that there is little difference in performance among the widely used kernels. See Wand and Jones (1994). If $\tau_{\mathrm{thresh}}$ is to be estimated with local linear regression using one of the seven popular kernel choices exhibited in Table 1, then the RDD has an asymptotic MSE that is about 2.3 times as large as that of the TBD, and the TBD will

TABLE 1

*An asymptotic comparison of the regression discontinuity designs with tie-breaker designs in Kernel regression-based estimation of $\tau_{\text{thresh}}$. The fourth column gives the quantity $((\tilde{C}_1\tilde{C}_2^4)/(C_1C_2^4))^{1/5}$, which is computed using (6) and (11) and rounded to 2 decimal places. The last column gives the quantity $\theta_*$ given by (20) rounded to 3 decimal places.*

| Kernel | Function $K(u)$ | Support | Relative AMSE | $\theta_*$ |
|---|---|---|---|---|
| Boxcar | $1/2$ | $[-1,1]$ | 2.30 | 0.354 |
| Triangular | $(1-\lvert u\rvert)_+$ | $[-1,1]$ | 2.27 | 0.359 |
| Epanechnikov | $\frac{3}{4}(1-u^2)_+$ | $[-1,1]$ | 2.31 | 0.351 |
| Quartic | $\frac{15}{16}(1-u^2)_+^2$ | $[-1,1]$ | 2.29 | 0.355 |
| Triweight | $\frac{35}{32}(1-u^2)_+^3$ | $[-1,1]$ | 2.28 | 0.357 |
| Tricube | $\frac{70}{81}(1-\lvert u\rvert^3)_+^3$ | $[-1,1]$ | 2.31 | 0.352 |
| Cosine | $\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)$ | $[-1,1]$ | 2.31 | 0.352 |

require 64 to 65 percent fewer samples than the RDD in order to achieve the same asymptotic MSE.

We think that a version of Theorem 1 will hold also for unbounded kernels such as the $\mathcal{N}(0,1)$ density under reasonable but stronger regularity conditions on $f(\cdot)$, $\mu_\pm(\cdot)$ and $\sigma_\pm(\cdot)$. We do not develop such a result as Cattaneo and Titiunik (2022) state that kernels with unbounded support are not used in RDD analysis.

## 5. Variance comparisons at fixed $h > \Delta$

The AMSE comparison in Section 4 depends upon the optimal TBD bandwidth, $\hat{h}_{\text{opt,TBD}}$, eventually becoming smaller than the positive experimental radius $\Delta$. However, the optimal $h$ converges to zero only at the very slow rate $N^{-1/5}$. Furthermore, the constant in that rate includes the factor $|\mu_+^{(2)}(t) - \mu_-^{(2)}(t)|^{-2/5}$ which could be very large. We believe that in many applied settings the optimal value of $h$ will not be smaller than $\Delta$. Then $\Delta/h$ is not necessarily within the support of the kernel and $Y$ values data from outside the experimental region are included in the local linear regression.

In this section, we complement the prior analysis with one where $h$ is fixed and larger than $\Delta$. We assume a symmetric kernel function that is Lipschitz continuous on its support.

The kernel regression estimate of $\hat{\tau}_{\text{thresh}}$ has a leading bias of $O(h^2)$. In the regime where the bandwidth is bigger than $\Delta$, mean squared optimality analysis for estimating $\tau_{\text{thresh}}$ is complicated by the fact that for the TBD there will often exist an $h > \Delta$ such that the constant in this $O(h^2)$ term vanishes. Remarkably, such a bandwidth depends only on the experimental radius $\Delta$ and the kernel $K$. It does not depend on $\mu_+$, $\mu_-$, $f$, or $N$. In Appendix D, we prove that under certain regularity conditions on $f$, $\mu_\pm$, and $K$, a bandwidth $h$ that solves $\nu_2^2 = 4\int_{\Delta/h}^\infty uK(u)\,\mathrm{d}u \int_{\Delta/h}^\infty u^3K(u)\,\mathrm{d}u$ removes the leading order bias, and moreover,

such a solution exists. See Table D.1 for numerical solutions of this equation for the kernel choices considered previously. We find that for these kernel choices, the bandwidth removing the leading order bias ranges from approximately $3.13\Delta$ for the Boxcar kernel to approximately $4.84\Delta$ for the Triweight kernel. We caution investigators against picking this bandwidth because it does not shrink with $N$. It could place too little weight on reducing variance for small $N$ and the third order bias term will be $O(1)$.

Due to the existence of a fixed bandwidth bigger than $\Delta$ that removes the leading order bias of $\hat{\tau}_{\text{thresh}}$, analysis of bias and mean squared optimality using second order Taylor expansions of $\mu_{\pm}(\cdot)$ would be misleading. Hence, we do not conduct an analysis similar to that seen in Section 4 for the regime where $h > \Delta$. For that regime, we instead restrict our attention to the variance in estimating $\tau_{\text{thresh}}$ at a fixed bandwidth $h$.

The variance of the local linear estimator $\hat{\tau}_{\text{thresh}}$ given in (4) and (5) can be computed as follows. The design matrix for the regression is $\mathcal{X} \in \mathbb{R}^{N\times 4}$ with $i$'th row $(1, x_i, Z_i, x_i Z_i)$. The response is $\mathcal{Y} = (Y_1, \ldots, Y_N)^{\mathsf{T}}$. For simplicity we assume $\text{var}(\mathcal{Y}\,|\,\mathcal{X}) = \sigma^2 I_N$ and without loss of generality we assume $t = 0$. The kernel weights are $K(x_i/h)$, and we let $\mathcal{W} = \mathcal{W}(h) \in \mathbb{R}^{N\times N} = \text{diag}(K(x_i/h))$. Then

$$\hat{\beta} = \hat{\beta}(\Delta) = (\mathcal{X}^{\mathsf{T}}\mathcal{W}\mathcal{X})^{-1}\mathcal{X}^{\mathsf{T}}\mathcal{W}\mathcal{Y}, \tag{21}$$

and under the assumption that $\text{var}(\mathcal{Y}\,|\,\mathcal{X}) = \sigma^2 I_N$ we have

$$\text{var}(\hat{\beta}\,|\,\mathcal{X}; \Delta) = (\mathcal{X}^{\mathsf{T}}\mathcal{W}\mathcal{X})^{-1}\mathcal{X}^{\mathsf{T}}\mathcal{W}^2\mathcal{X}(\mathcal{X}^{\mathsf{T}}\mathcal{W}\mathcal{X})^{-1}\sigma^2. \tag{22}$$

Formula (21) for $\hat{\beta}$ matches the familiar generalized least squares formula for the case where $\text{var}(\mathcal{Y}\,|\,\mathcal{X}) = \mathcal{W}\sigma^2$. Here $\mathcal{W}$ arises from weights that are not of inverse variance type and hence the formula for $\text{var}(\hat{\beta}\,|\,\mathcal{X}; \Delta)$ involves a $\mathcal{W}^2$ factor and less cancellation than we might have expected. The boxcar kernel is special because then $K(x_i/h) \in \{0, 1\}$ equals its own square. In that case $\text{var}(\hat{\beta}\,|\,\mathcal{X}; \Delta) = (\mathcal{X}^{\mathsf{T}}\mathcal{W}\mathcal{X})^{-1}\sigma^2$. The estimator is $\hat{\tau}_{\text{thresh}} = 2\hat{\beta}_3$. Therefore, we study $\text{var}(\hat{\beta}_3\,|\,\mathcal{X}; \Delta)$ under a tie-breaker design as $(\text{var}(\hat{\beta}\,|\,\mathcal{X}; \Delta))_{3,3}$ using the expression in (22).

At the stage where the experiment is being designed and $\Delta$ is being chosen, the investigator does not have much information about $\mathcal{X} \in \mathbb{R}^{N\times 4}$ but we will later see, quite a bit is known about the quantity $N \times \text{var}(\hat{\beta}_3\,|\,\mathcal{X}; \Delta)/\sigma^2$. For $x_i$ from a real dataset, we see in Section 6 (e.g. Figure 5) that this quantity does not vary much for different simulations of the random treatment assignments $(Z_i)_{i=1}^N$. To get theoretical insight, we turn our attention to the uniformly spaced setting with $x_i = (2i - N - 1)/N$ to develop tractable theoretical results. We give an asymptotic justification for this assumption using results from Fan and Gijbels (1996) in Section 6. This rank transformation is also used in Owen and Varian (2020).

For $x_i = (2i - N - 1)/N$, the matrices $\mathcal{X}^{\mathsf{T}}\mathcal{W}\mathcal{X}/N$ and $\mathcal{X}^{\mathsf{T}}\mathcal{W}^2\mathcal{X}/N$ contain

elements that can be approximated by integrals of the form

$$\mathcal{I}^{rst} = \mathcal{I}^{rst}(\Delta, h, K) \equiv \frac{1}{2}\int_{-1}^{1} x^r \mathbb{E}(Z^s \,|\, x; \Delta) K\Big(\frac{x}{h}\Big)^t \,\mathrm{d}x \tag{23}$$

for integer exponents $r$, $s$ and $t$. Our expressions will simplify somewhat because $Z^2 = 1$ making every $\mathcal{I}^{r,2,t} = \mathcal{I}^{r,0,t}$ and also because both $x$ and $\mathbb{E}(Z \mid x; \Delta)$ are antisymmetric functions of $x$ making them orthogonal to $K(x/h)$ which we have assumed to be symmetric. The error in those moment approximations is $O_p(N^{-1/2})$ if the $Z_i$ are independent random variables. The error can be much less with other sampling schemes. For instance, we could use stratified sampling, forming pairs of subjects $(i, i+1)$ in the experimental region and randomly setting $Z_i = \pm 1$ and $Z_{i+1} = -Z_i$. We will use $\approx$ to describe approximations that are $O_p(N^{-1/2})$ or better.

Applying first $Z^2 = 1$ and then using symmetry and anti-symmetry

$$\frac{1}{N}\mathcal{X}^{\mathsf{T}}\mathcal{W}\mathcal{X} \approx \begin{array}{c} \\ 1 \\ x \\ z \\ xz \end{array}\!\!\begin{array}{c} \begin{array}{cccc} 1 & x & z & xz \end{array} \\ \begin{bmatrix} \mathcal{I}^{001} & \mathcal{I}^{101} & \mathcal{I}^{011} & \mathcal{I}^{111} \\ \mathcal{I}^{101} & \mathcal{I}^{201} & \mathcal{I}^{111} & \mathcal{I}^{211} \\ \mathcal{I}^{011} & \mathcal{I}^{111} & \mathcal{I}^{021} & \mathcal{I}^{121} \\ \mathcal{I}^{111} & \mathcal{I}^{211} & \mathcal{I}^{121} & \mathcal{I}^{221} \end{bmatrix} \end{array} = \begin{bmatrix} \mathcal{I}^{001} & 0 & 0 & \mathcal{I}^{111} \\ 0 & \mathcal{I}^{201} & \mathcal{I}^{111} & 0 \\ 0 & \mathcal{I}^{111} & \mathcal{I}^{001} & 0 \\ \mathcal{I}^{111} & 0 & 0 & \mathcal{I}^{201} \end{bmatrix}.$$

Because $K^2(\cdot)$ is also a symmetric function we also get

$$\frac{1}{N}\mathcal{X}^{\mathsf{T}}\mathcal{W}^2\mathcal{X} \approx \begin{bmatrix} \mathcal{I}^{002} & 0 & 0 & \mathcal{I}^{112} \\ 0 & \mathcal{I}^{202} & \mathcal{I}^{112} & 0 \\ 0 & \mathcal{I}^{112} & \mathcal{I}^{002} & 0 \\ \mathcal{I}^{112} & 0 & 0 & \mathcal{I}^{202} \end{bmatrix}.$$

From all of the symmetries involved in the 32 components of these two matrices, we need to consider at most six distinct integrals. We rewrite those matrices, beginning with

$$\frac{1}{N}\mathcal{X}^{T}\mathcal{W}\mathcal{X} \approx \begin{bmatrix} \kappa_0 & 0 & 0 & \phi(\Delta) \\ 0 & \kappa_2 & \phi(\Delta) & 0 \\ 0 & \phi(\Delta) & \kappa_0 & 0 \\ \phi(\Delta) & 0 & 0 & \kappa_2 \end{bmatrix} \tag{24}$$

where

$$\begin{aligned} \kappa_0 &= \frac{1}{2}\int_{-1}^{1} K\Big(\frac{x}{h}\Big)\,\mathrm{d}x, \quad \kappa_2 = \frac{1}{2}\int_{-1}^{1} x^2 K\Big(\frac{x}{h}\Big)\,\mathrm{d}x, \quad \text{and} \\ \phi(\Delta) &= \frac{1}{2}\int_{-1}^{-\Delta} (-x) K\Big(\frac{x}{h}\Big)\,\mathrm{d}x + \frac{1}{2}\int_{\Delta}^{1} x K\Big(\frac{x}{h}\Big)\,\mathrm{d}x = \int_{\Delta}^{1} x K\Big(\frac{x}{h}\Big)\,\mathrm{d}x. \end{aligned} \tag{25}$$

Note that $\kappa_0$ and $\kappa_2$ may depend on $h$ but they do not depend on $\Delta$. A similar argument shows that

$$\frac{1}{N}\mathcal{X}^T\mathcal{W}^2\mathcal{X} \approx \begin{bmatrix} \lambda_0 & 0 & 0 & \psi(\Delta) \\ 0 & \lambda_2 & \psi(\Delta) & 0 \\ 0 & \psi(\Delta) & \lambda_0 & 0 \\ \psi(\Delta) & 0 & 0 & \lambda_2 \end{bmatrix} \tag{26}$$

for

$$\begin{aligned} \lambda_0 = \frac{1}{2}\int_{-1}^{1} K^2\Big(\frac{x}{h}\Big)\,\mathrm{d}x, \quad \lambda_2 = \frac{1}{2}\int_{-1}^{1} x^2K^2\Big(\frac{x}{h}\Big)\,\mathrm{d}x, \quad \text{and} \\ \psi(\Delta) = \int_{\Delta}^{1} xK^2\Big(\frac{x}{h}\Big)\,\mathrm{d}x. \end{aligned} \tag{27}$$

Now we are ready to describe the asymptotic variance of $\hat{\beta}_3$.

**Theorem 2.** *Let $x_i = (2i - N - 1)/N$, select $Z_i \in \{-1, 1\}$ by the tie-breaker equation (1) with $t = 0$. Let $Y_i$ be uncorrelated random variables with common variance $\sigma^2$, conditionally on $\mathcal{X} = ((1, x_1, Z_1, x_1Z_1), \cdots, (1, x_N, Z_N, x_NZ_N))$. Next, for a symmetric kernel $K(\cdot) \geqslant 0$ that is Lipschitz continuous on its support and a bandwidth $h > 0$, let $\hat{\beta}$ be estimated by the kernel weighted regression (4). Then*

$$N\mathrm{var}(\hat{\beta}_3 \,|\, \mathcal{X}; \Delta) = \frac{\sigma^2\big(\kappa_2^2\lambda_0 - 2\kappa_2\phi(\Delta)\psi(\Delta) + \lambda_2\phi^2(\Delta)\big)}{\big(\kappa_0\kappa_2 - \phi^2(\Delta)\big)^2} + O_p\Big(\frac{1}{\sqrt{N}}\Big), \tag{28}$$

*where $\kappa_0$, $\kappa_2$ and $\phi(\Delta)$ are defined in (25) and $\lambda_0$, $\lambda_2$ and $\psi(\Delta)$ are defined in (27).*

*Proof.* Reordering the components of $\beta$ we find after substituting equations (24) and (26) into (22) that $\sqrt{N}(\hat{\beta}_1, \hat{\beta}_4, \hat{\beta}_2, \hat{\beta}_3)$ has variance

$$\begin{pmatrix} \kappa_0 & \phi & 0 & 0 \\ \phi & \kappa_2 & 0 & 0 \\ 0 & 0 & \kappa_0 & \phi \\ 0 & 0 & \phi & \kappa_2 \end{pmatrix}^{-1} \begin{pmatrix} \lambda_0 & \psi & 0 & 0 \\ \psi & \lambda_2 & 0 & 0 \\ 0 & 0 & \lambda_0 & \psi \\ 0 & 0 & \psi & \lambda_2 \end{pmatrix} \begin{pmatrix} \kappa_0 & \phi & 0 & 0 \\ \phi & \kappa_2 & 0 & 0 \\ 0 & 0 & \kappa_0 & \phi \\ 0 & 0 & \phi & \kappa_2 \end{pmatrix}^{-1} \sigma^2 + O_p\Big(\frac{1}{\sqrt{N}}\Big).$$

Now (28) follows directly by matrix inversion and multiplication. □

Our Lipschitz condition on the kernel $K$ is present for technical reasons. Using a formulation with fixed and discrete $x_i = (2i - N - 1)/N$, this condition allows us to obtain the same error rate of $O_p(N^{-1/2})$ as would be obtained using the formulation with random $x_i \overset{\text{IID}}{\sim} \mathbb{U}[-1, 1]$. Without the Lipschitz condition, an adversarially chosen kernel $K(\cdot)$ might have point discontinuities at every rational multiple of the bandwidth $h$. We remark that our Lipschitz condition can be loosened to a 1/2-Hölder continuity condition, with details available from the first author upon request.

The variance formula in Theorem 2 does not require the linear model (2) to hold. When it does not hold there will generally be some bias where $\mathbb{E}(2\hat{\beta}_3 \mid \mathcal{X}; \Delta) \neq \mu_+(0) - \mu_-(0)$. We suppose that the user will choose an $h$ to appropriately navigate the bias variance tradeoff, but that step takes place after the outcomes $Y_i$ are observed, which are not available when $\Delta$ is chosen, so we resort to comparing the variance for any choice of $h$.

We are primarily interested in comparing the asymptotic variance of $\hat{\tau}_0 = 2\hat{\beta}_3$ for various choices of $\Delta$. We especially want to compare the efficiency of tie-breaker designs with $\Delta > 0$ to the RDD with $\Delta = 0$. To do this we consider the efficiency ratio

$$\mathrm{Eff}^{(N)}(\Delta) \equiv \frac{\mathrm{var}(\hat{\tau}_{\mathrm{thresh}} \,|\, \mathcal{X}; 0)}{\mathrm{var}(\hat{\tau}_{\mathrm{thresh}} \,|\, \mathcal{X}; \Delta)} = \frac{\mathrm{var}(\hat{\beta}_3 \,|\, \mathcal{X}; 0)}{\mathrm{var}(\hat{\beta}_3 \,|\, \mathcal{X}; \Delta)}. \tag{29}$$

Using Theorem 2, $\mathrm{Eff}^{(N)}(\Delta)$ converges in probability to the asymptotic efficiency ratio

$$\mathrm{Eff}(\Delta) = \frac{\big(\kappa_2^2\lambda_0 - 2\kappa_2\phi(0)\psi(0) + \lambda_2\phi^2(0)\big)\big(\kappa_0\kappa_2 - \phi^2(\Delta)\big)^2}{\big(\kappa_2^2\lambda_0 - 2\kappa_2\phi(\Delta)\psi(\Delta) + \lambda_2\phi^2(\Delta)\big)\big(\kappa_0\kappa_2 - \phi^2(0)\big)^2} \tag{30}$$

using quantities that we defined at (25) and (27).

### *Efficiency with boxcar and triangular kernels*

In this subsection we present the efficiency ratios under the conditions of Theorem 2 for the two kernels of greatest interest: the boxcar kernel and the triangular kernel. We work with $x_i = (2i - N - 1)/N$ throughout this subsection.

For the boxcar kernel $K_{\mathrm{BC}}(u) = 1_{|u| \leqslant 1}$, we can assume without loss of generality that $h \leqslant 1$ because there are no data with $|x_i - t| = |x_i| > 1$, and then any $h > 1$ will give the same estimate as $h = 1$. We find for this kernel that

$$\kappa_0 = \lambda_0 = h, \quad \kappa_2 = \lambda_2 = \frac{h^3}{3}, \quad \text{and} \quad \phi(\Delta) = \psi(\Delta) = \frac{(h^2 - \Delta^2)_+}{2}. \tag{31}$$

Using some foresight, we define the local tie-breaker constant $\delta = \Delta/h$. This is the fraction of the local regression region in which the treatment was assigned at random.

**Proposition 1.** *Under the conditions of Theorem 2 and using the boxcar kernel $K_{\mathrm{BC}}$, the asymptotic efficiency ratio of the tie-breaker design is*

$$\mathrm{Eff}_{\mathrm{BC}} = 1 + 6\delta^2 - 3\delta^4 \tag{32}$$

*for $\delta = \Delta/h \leqslant 1$. If $\delta > 1$, then $\mathrm{Eff}_{\mathrm{BC}} = 4$.*

*Proof.* Because many quantities from (31) are identical, substituting them into (30) produces numerous simplifications that yield

$$\mathrm{Eff}_{\mathrm{BC}} = \frac{\kappa_0\kappa_2 - \phi^2(\Delta)}{\kappa_0\kappa_2 - \phi^2(0)} = \frac{\frac{h^4}{3} - \frac{(h^2-\Delta^2)_+^2}{4}}{\frac{h^4}{3} - \frac{h^4}{4}} = 4 - 3(1-\delta^2)_+^2.$$

For $0 \leqslant \delta < 1$ formula (32) follows from expanding the quadratic while for $\delta > 1$ the positive part term vanishes. □

Choosing $h = 1$ makes the local regression a global one. We then get the same efficiency ratio as in equation (6) from Owen and Varian (2020). By taking derivatives it is easy to show that the efficiency ratio in (32) is strictly increasing as the local amount of experimentation $\delta$ varies over the interval $0 < \delta < 1$. Figure 2 plots $\mathrm{Eff}_{\mathrm{BC}}$ versus $\delta$.

The triangular spike kernel $K_{\mathrm{TS}}(x) = (1 - |x|)_+$ (triangular kernel for short) is more complicated than the boxcar kernel because for it, $K^2$ is not proportional to $K$. Once again, we assume that $h \in [0, 1]$. For this kernel we compute

$$\kappa_0 = \frac{h}{2}, \quad \kappa_2 = \frac{h^3}{12}, \quad \lambda_0 = \frac{h}{3}, \quad \text{and} \quad \lambda_2 = \frac{h^3}{30}$$

and then using $\delta = \Delta/h$, we get

$$\phi(\Delta) = \frac{h^2}{6}(1 - 3\delta^2 + 2\delta^3) \quad \text{and} \quad \psi(\Delta) = \frac{h^2}{12}(1 - 6\delta^2 + 8\delta^3 - 3\delta^4).$$

**Proposition 2.** *Under the conditions of Theorem 2 and using the triangular kernel $K_{\mathrm{TS}}$, the asymptotic efficiency of the tie-breaker design is*

$$\mathrm{Eff}_{\mathrm{TS}} = \frac{2\big(3 - 2(1 - 3\delta^2 + 2\delta^3)^2\big)^2}{5 - 5(1 - 3\delta^2 + 2\delta^3)(1 - 6\delta^2 + 8\delta^3 - 3\delta^4) + 2(1 - 3\delta^2 + 2\delta^3)^2} \tag{33}$$

*for $\delta = \Delta/h \leqslant 1$.*

*Proof.* This follows from plugging in the values of $\kappa_0$, $\kappa_2$, $\lambda_0$, $\lambda_2$, $\phi(\Delta)$, and $\psi(\Delta)$ for the triangular kernel into (30). See Appendix E for the explicit calculations. □

The second panel in Figure 2 shows $\mathrm{Eff}_{\mathrm{TS}}$ versus the local experiment size $\delta$. The efficiency curve has a similar monotone increasing shape as we saw for the boxcar kernel. The maximum efficiency ratio, at $\delta = 1$, is $18/5 = 3.6$ instead of 4. The efficiency ratio is a rational function of $\delta$ with a numerator of degree 12 and a denominator of degree 7. It is strictly increasing on the interval $0 < \delta < 1$, though the proof is lengthy enough to move to the Appendix.

**Proposition 3.** *The derivative of $\mathrm{Eff}_{\mathrm{TS}}$ with respect to $\delta$ is positive for $0 < \delta < 1$.*

*Proof.* See Appendix F. □

## 6. Classroom size data

We explored the efficiency ratio for the tie-breaker design for $x_i$ with a uniform distribution. While that can be arranged by using ranks, in other situations we might prefer to use the original value of a running variable and those might

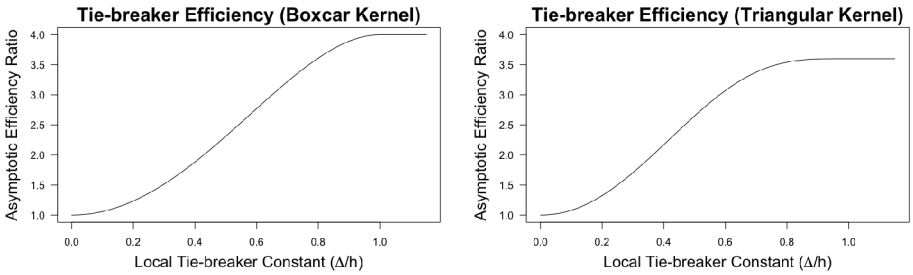



**Fig 2:** The left panel shows the efficiency ratio of the tie-breaker design for uniform $x_i$ and the boxcar kernel as a function of $\delta = \Delta/h$. The right panel shows this efficiency ratio for the triangular kernel.

not be uniformly distributed. We show how to do this using a dataset from Angrist and Lavy (1999) on classroom sizes.

Angrist and Lavy (1999) studied the causal effect of classroom size on test performance of elementary school students in Israel. In Israel, the Maimonides rule mandates that elementary school classes cannot exceed 40 students. If a school has 41 students enrolled in a particular grade that grade must be split into two classes. Note that grades that have 40 or fewer enrolled students are allowed to split into multiple classes and that grades with slightly more than 40 students occasionally violate the Maimonides rule and do not split into multiple classes. Despite this, we can consider this a setting for RDD where the treatment variable is whether or not the school is legally mandated to split a particular grade into smaller classes.

The dataset, published on the Harvard Dataverse (Angrist and Lavy, 2009), has verbal and math scores for 3rd, 4th and 5th graders across Israel. We chose to focus exclusively on 4th grade verbal scores as our response variable and 4th grade enrollments as our assignment variable because Angrist and Lavy (1999) suggest that a slightly significant effect of the treatment on 4th grade verbal scores exists. Even though the data were not generated by a tie-breaker we can still compute the relative efficiency that a tie-breaker design would have had.

To simplify the analysis, we removed all schools that either had more than 80 students or more than two 4th-grade classes from the dataset. We further removed all schools that had NA entries for either class size or verbal scores, leaving $N = 711$ schools in our filtered dataset. See Figure 3 for a visualization of the distribution of the 4th grade enrollments and Figure 4 for visualizations of the local linear regression based-RDD on this dataset using boxcar and triangular kernels. We use the bandwidths $h_{\mathrm{IK}}$ given by the Imbens and Kalyanaraman (2012) procedure, which were computed using that paper's MATLAB code. The apparent benefit from smaller classrooms is positive but small and it turns out, not statistically significant in this analysis. The 95% confidence interval (assuming homoscedastic errors) for the effect size at the boundary of the local linear regression-based RDD was $(-1.5, 9.2)$ when a boxcar kernel with band-

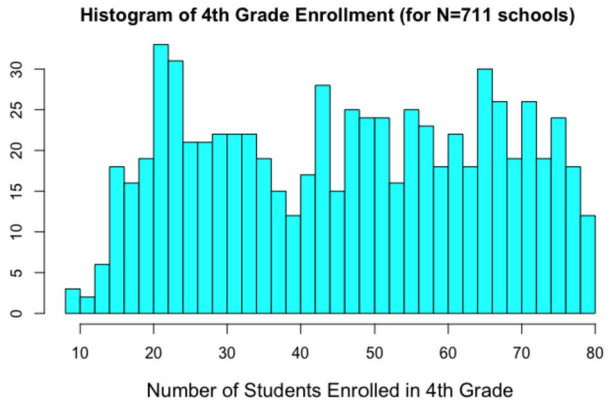



**Fig 3:** A histogram of 4th grade enrollments for our filtered dataset (with schools exceeding 80 4th grade students or three 4th grade classes removed).

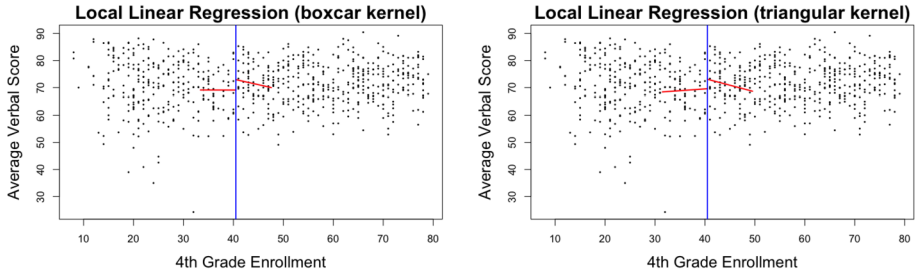



**Fig 4:** RDD fit to the 4th grader verbal scores from the Angrist and Lavy (2009) dataset when using a boxcar kernel (left) and a triangular kernel (right). For these two fits, the bandwidths $h_{\mathrm{IK,BC}}$ and $h_{\mathrm{IK,TS}}$ were chosen as in Imbens and Kalyanaraman (2012).

width $h_{\mathrm{IK,BC}} = 7.09$ was used. The 95% confidence interval for the effect size at boundary of this RDD was $(-2.4, 9.4)$ when a triangular kernel with bandwidth $h_{\mathrm{IK,TS}} = 9.02$ was used.

Next we illustrate how an investigator can estimate the efficiency ratio of tie-breaker designs as a function of $\Delta$ on sample values of the assignment variable. First we translate the data, replacing $x_i$ by $x_i - 40.5$ to move the threshold from $t = 40.5$ to $t = 0$. Next, for each $\Delta$ of interest we use 1000 Monte Carlo samples to estimate $\mathrm{var}(\hat{\beta}_3 \mid \mathcal{X}; \Delta)$ and also $\mathrm{var}(\hat{\beta}_3 \mid \mathcal{X}; 0)$, both up to a constant $\sigma^2$. That gives us 1000 efficiency ratios $\mathrm{Eff}^{(N)}(\Delta) = \mathrm{var}(\hat{\beta}_3 \,|\, \mathcal{X}; 0)/\mathrm{var}(\hat{\beta}_3 \,|\, \mathcal{X}; \Delta)$ for each $\Delta$. In each of our 1000 samples, we simulate random assignments $Z_i$ for a tie-breaker design at the given experimental radius $\Delta$. The random assignments are stratified: in each consecutive pair of classroom sizes in the experimental region, one was randomly chosen to have $Z = 1$ and the other got $Z = -1$. The $x_i$ and the random $Z_i$ let us compute the matrices $\mathcal{X}$ and $\mathcal{W}$ defined in the

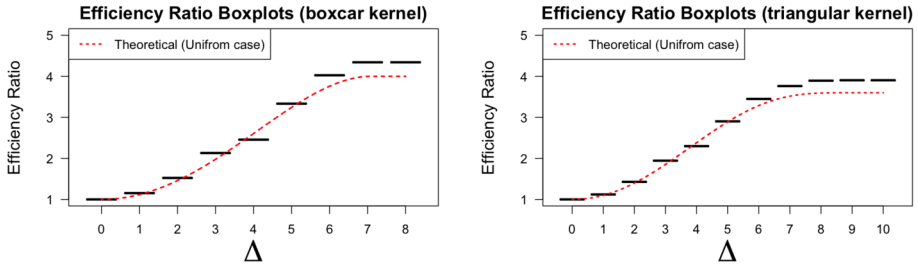



**Fig 5:** Boxplots of the Monte-Carlo efficiency ratio estimates for various values of $\Delta \in \mathbb{N}$ when using a boxcar kernel (left) and a triangular kernel (right). For both kernels we used bandwidths from Figure 4: $h_{\mathrm{IK,BC}} = 7.09$ and $h_{\mathrm{IK,TS}} = 9.02$. The dashed lines give the theoretical efficiency curves under the assumption of a uniform assignment variable, given by Propositions 1 (left) and 2 (right).

beginning of Section 5, from which we compute a non-asymptotic $\mathrm{var}(\hat{\beta}_3 \,|\, \mathcal{X}; \Delta)$ using (22). We do not simulate any $Y_i$ values because efficiency only depends on $\mathcal{X}$, and we are retaining the bandwidths from the Imbens and Kalyanaraman (2012) procedure on the original data. A more detailed simulation randomizing the bandwidth choice is out of scope. Our simulations demonstrate that the TBD is more efficient at each fixed $h$, so we expect that it will also be more efficient at a randomly chosen $h$. There could be exceptions if the bandwidth is adversarially correlated with the estimation errors but we do not think that is likely.

Figure 5 shows boxplots of 1000 simulated $\mathrm{Eff}^{(N)}(\Delta)$ values for various choices of $\Delta \in \mathbb{N}$ to plot the full efficiency curve. It is clear from Figure 5 that with stratified allocations the efficiency is very reproducible. Figure 6 shows results for different bandwidths, ranging from $h_{\mathrm{IK}}/2$ to $3h_{\mathrm{IK}}/2$. Because the efficiencies are so reproducible given the bandwidth, we just plot curves of the mean and standard deviations of estimated Eff values. For both the boxcar and triangular kernels, we see that the tie-breaker design is reproducibly more efficient than the RDD and the effect increases as $\delta = \Delta/h$ increases for all $h$ we studied. The efficiency curves for this dataset under various bandwidth choices look similar to the theoretical efficiency curves derived in Section 5 for the case of a uniform assignment variable.

For a further discussion of the Maimonides rule, see Angrist et al. (2019). They consider different data sets and also investigate the possibility that the class sizes are sometimes manipulated to be above the threshold triggering a classroom split.

### *Comparison with theoretical results for uniform assignment variable*

Our theoretical analysis in Section 5 is for a uniformly spaced assignment variable. We can offer one explanation for why the empirical efficiencies on non-uniformly distributed data look so similar to the theoretical ones for uniformly

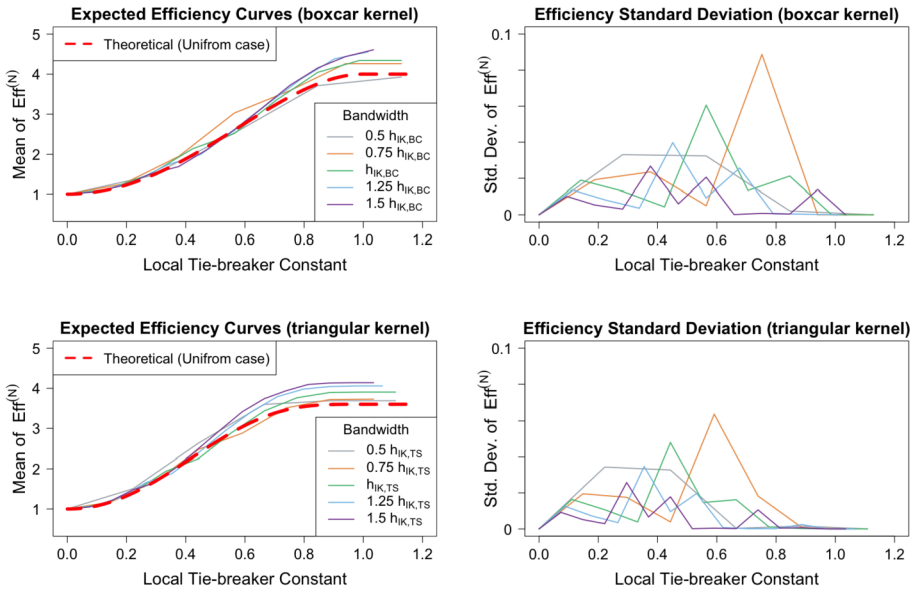



**Fig 6:** Monte-Carlo estimates of the expected value (left) and standard deviation (right) of $\mathrm{Eff}^{(N)}(\Delta)$ versus $\Delta/h$ for the Angrist and Lavy (2009) dataset of 4th grader verbal scores. For these plots a boxcar kernel (top) and a triangular kernel (bottom) were used. The bandwidths plotted are scalar multiples of $h_{\mathrm{IK,BC}}$ and $h_{\mathrm{IK,TS}}$ from the procedure of Imbens and Kalyanaraman (2012). The legend for the plots on the right is the same as for the plots on the left. These curves are not smooth because, to avoid redundancy, only points that corresponded to integer values of $\Delta$ were used. The thick dashed lines give the theoretical efficiency curves $\mathrm{Eff}_{\mathrm{BC}}$ (top left) and $\mathrm{Eff}_{\mathrm{TS}}$ (bottom left) derived in Section 5.

distributed data (see the left panels in Figure 6). The explanation uses some results about non-parametric regression from Fan and Gijbels (1996, Table 2.1). Nonparametric regression estimates $\hat{\mu}(t)$ typically have an asymptotic variance where the leading term is proportional to $1/f(t)$ where $f$ is the probability density of the $x_i$. This arises because the local sample size is asymptotically proportional to $f(t)$. Hence, when considering nonuniform distributions, the $1/f(t)$ factors in the leading order variance terms will cancel out when computing the efficiency ratios. Some of the nonparametric regression estimators, such as the Nadaraya-Watson estimator, have a lead term in their bias that depends on the derivative $f'(t)$, and while $f'(t) = 0$ for uniformly distributed data, it is not zero in general. Kernel weighted least squares methods (with symmetric $K(\cdot)$) do not have a dependency on $f'(t)$ in their bias. There is a curvature bias from $\mu''(t)$ but that is not related to the sampling distribution of the $x_i$. The lead terms in bias and variance for local linear regressions do not distinguish between distributions with the same value of $f(t)$ but different $f'(t)$. Thus the effects of non-uniformity of $X$ are asymptotically negligible.

## 7. Discussion

If an investigator is able to implement a 3-level tie-breaker design with any experimental radius $\Delta \in (0, \Delta_{\max})$, our results show that the TBD has considerable statistical advantages over the RDD.

The most obvious advantage is that the TBD allows estimation of multiple causal parameters of interest including the average treatment effect over subjects with $x \in (t-\Delta, t+\Delta)$ as well as the expected treatment effect at any particular $x \in (t-\Delta, t+\Delta)$. The former is estimable at a faster rate and with fewer assumptions, whereas the latter may still be of interest for choosing a future policy threshold. Meanwhile, the RDD only allows estimation of $\tau_{\text{thresh}}$, the expected treatment effect at $x = t$.

Even if the only goal is estimation of $\tau_{\text{thresh}}$, our results indicate a statistical advantage to running a TBD rather than an RDD and an advantage to picking a larger experimental radius $\Delta \in (0, \Delta_{\max})$. As seen in Section 4, to achieve the same asymptotic MSE in mean squared optimal estimation of $\tau_{\text{thresh}}$, a TBD would require roughly 64 percent fewer samples than would be needed for an RDD. Moreover, the asymptotic advantage for a TBD is largely driven by its lower variance (Figure 1). Hence, if the convenient, but controversial, method of undersmoothing to construct asymptotically valid confidence intervals for $\tau_{\text{thresh}}$ is used instead of more nearly optimal approaches, the TBD would exhibit even greater advantages over the RDD. We point readers to the introduction of Calonico, Cattaneo and Farrell (2018) for an overview of the history of undersmoothing, and Calonico, Cattaneo and Farrell (2019) for a modern approach to constructing confidence intervals that has better coverage properties than undersmoothing has.

In terms of the statistical advantages of picking a larger $\Delta$, Owen and Varian (2020) found an efficiency advantage for the tie-breaker in a global regression, wherein the estimation variance decreased monotonically in $\Delta$. We provide a comparable finding for the now more standard local linear regression approach: for any fixed bandwidth $h$, we see a theoretical efficiency that increases with the amount $\Delta$ of experimentation. We have not investigated the effect of $\Delta$ on the subsequent choice of $h$ when $\hat{h}_{\text{opt,TBD}} > \Delta$, although one candidate choice is an $h > \Delta$ that removes the leading order bias term, which we derived in Appendix D.

There is room for an improved estimator of $\gamma$ in the TBD context which uses data from both treatments on both sides of the threshold $t$. We leave this for further work. A critical ingredient is the estimation of $\mu_{\pm}^{(2)}(t)$. Compared to the method in Imbens and Kalyanaraman (2012), one could use a bandwidth tuned for an internal point $t$ instead of one tuned for an endpoint. Also the curvature estimates in Imbens and Kalyanaraman (2012) use local quadratic regressions while Fan and Gijbels (1996, p 63) suggest using local cubic regressions for curvature estimation at an interior point.

## Appendix A: Proof of Lemma 2

Without loss of generality suppose $t = 0$ (if this is not the case we can define a new assignment variable to be a translation of the original assignment variable by $t$). It is convenient to define

$$\mathcal{Z}_+ \equiv \{i \in \{1, 2, \ldots, N\} \,:\, Z_i = 1\} \quad \text{and} \quad \mathcal{Z}_- \equiv \{i \in \{1, 2, \ldots, N\} : \, Z_i = -1\}.$$

Also define $N_\pm \equiv |\mathcal{Z}_\pm|$. Letting $K_h(u) \equiv K(u/h)/h$, define

$$F_{j,\pm} \equiv \frac{1}{N_\pm} \sum_{i \in \mathcal{Z}_\pm} X_i^j K_h(X_i) \quad \text{and} \quad G_{j,\pm} \equiv \frac{1}{N_\pm} \sum_{i \in \mathcal{Z}_\pm} X_i^j K_h^2(X_i) \sigma_\pm^2(X_i),$$

for nonnegative integers $j$.

Next, let $f_\pm$ be the conditional density of $X_i$ given that $Z_i = \pm 1$. That is

$$f_+(x) = \begin{cases} \frac{f(x)}{\Pr(Z_i=1)}, & x \geqslant \Delta \\ \frac{f(x)/2}{\Pr(Z_i=1)}, & |x| < \Delta \\ 0, & x \leqslant -\Delta \end{cases} \quad \text{and} \quad f_-(x) = \begin{cases} 0, & x \geqslant \Delta \\ \frac{f(x)/2}{\Pr(Z_i=-1)}, & |x| < \Delta \\ \frac{f(x)}{\Pr(Z_i=-1)}, & x \leqslant -\Delta. \end{cases} \tag{34}$$

We now prove some helpful small bandwidth approximations for $F_{j,\pm}$ and $G_{j,\pm}$.

**Lemma A.1.** *In the setting of Lemma 2, for nonnegative integers $j$,*

$$F_{j,\pm} = h^j f_\pm(0) \nu_j + o_p(h^j)$$

*with $\nu_j = \int_{-\infty}^{\infty} u^j K(u) \, \mathrm{d}u$.*

*Proof.* We will prove this for $F_{j,+}$ and the proof for $F_{j,-}$ will be identical. First note that $F_{j,+}$ is the average of $N_+$ IID random variables so

$$F_{j,+} = \mathbb{E}[X^j K_h(X) | Z = 1] + O_p\Big(\sqrt{\mathrm{Var}[X^j K_h(X) | Z = 1]/N_+}\Big).$$

Now note that

$$\begin{aligned}
&\mathbb{E}[X^j K_h(X) | Z = 1] \\
&\quad = \int_{-\infty}^{\infty} \frac{1}{h} K(x/h) x^j f_+(x) \, \mathrm{d}x \\
&\quad = h^j \int_{-\infty}^{\infty} K(u) u^j f_+(uh) \, \mathrm{d}u \\
&\quad = h^j \int_{-\infty}^{\infty} K(u) u^j f_+(0) \, \mathrm{d}u + h^{j+1} \int_{-\infty}^{\infty} K(u) u^{j+1} \frac{f_+(hu) - f_+(0)}{hu} \, \mathrm{d}u \\
&\quad = h^j f_+(0) \nu_j + h^{j+1} \int_{-\infty}^{\infty} K(u) u^{j+1} \frac{f_+(hu) - f_+(0)}{hu} \, \mathrm{d}u \\
&\quad = h^j f_+(0) \nu_j + O(h^{j+1}),
\end{aligned}$$

where the last equality holds because $K(\cdot)$ is bounded with bounded support, $f_+$ is continuously differentiable at 0, and asymptotically $h \to 0$.

Meanwhile, the term $\mathrm{Var}[X^j K_h(X)|Z=1]$ is upper bounded by

$$\begin{aligned}\mathbb{E}[(K_h(X))^2 X^{2j}|Z=1] &= \frac{1}{h^2}\int_{-\infty}^{\infty}\big(K(x/h)\big)^2 x^{2j} f_+(x)\,\mathrm{d}x \\ &= h^{2j-1}\int_{-\infty}^{\infty} K^2(u)u^{2j} f_+(hu)\,\mathrm{d}u \\ &= O(h^{2j-1}),\end{aligned}$$

where the last equality holds because $K(\cdot)$ is bounded with bounded support, $f_+$ is continuous and strictly positive at 0, and asymptotically $h \to 0$. Thus

$$\begin{aligned}\frac{\mathrm{Var}[X^j K_h(X)|Z=1]}{N_+} &\leqslant \frac{\mathbb{E}[(K_h(X_i))^2 X_i^{2j}|Z=1]}{N_+} \\ &= \frac{O(h^{2j-1})}{N_+} = O\Big(\frac{h^{2j}}{hN}\Big)\cdot\frac{N}{N_+} = o_p(h^{2j}),\end{aligned}$$

where last equality holds because as $N \to \infty$, $hN \to \infty$, so $h^{2j}/hN = o(h^{2j})$, and because $N/N_+ = O_p(1)$.

Combining previous results

$$F_{j,+} = h^j f_+(0)\nu_j + O(h^{j+1}) + O_p(o_p(h^j)) = h^j f_+(0)\nu_j + o_p(h^j). \qquad \square$$

**Lemma A.2.** *In the setting of Lemma 2, for nonnegative integers $j$,*

$$G_{j,\pm} = h^{j-1}\sigma_\pm^2(0) f_\pm(0)\pi_j(1+o_p(1))$$

*with $\pi_j = \int_{-\infty}^{\infty} u^j K^2(u)\,\mathrm{d}u$.*

*Proof.* We will prove this for $G_{j,+}$ and the proof for $G_{j,-}$ will be identical. First note that $G_{j,+}$ is the average of $N_+$ IID random variables so

$$G_{j,+} = \mathbb{E}[X^j K_h^2(X)\sigma_+^2(X)|Z=1] + O_p\Big(\sqrt{\mathrm{Var}[X^j K_h^2(X)\sigma_+^2(X)|Z=1]/N_+}\Big).$$

Now note that $\mathbb{E}[X^j K_h^2(X)\sigma_+^2(X)|Z=1]$ equals

$$\begin{aligned}&\frac{1}{h^2}\int_{-\infty}^{\infty} K^2(x/h)x^j\sigma_+^2(x)f_+(x)\,\mathrm{d}x \\ &= h^{j-1}\int_{-\infty}^{\infty} K^2(u)u^j\sigma_+^2(hu)f_+(hu)\,\mathrm{d}u \\ &= h^{j-1}\pi_j\sigma_+^2(0)f_+(0) + h^{j-1}\int_{-\infty}^{\infty} K^2(u)u^j\big[\sigma_+^2(hu)f_+(hu) - \sigma_+^2(0)f_+(0)\big]\,\mathrm{d}u \\ &= h^{j-1}\pi_j\sigma_+^2(0)f_+(0) + o(h^{j-1}), \qquad (35)\end{aligned}$$

where the last equality holds because $K(\cdot)$ is bounded with bounded support, $\sigma_+^2(\cdot)f_+(\cdot)$ is continuous at 0, and asymptotically $h \to 0$.

Meanwhile, the term $\mathrm{Var}[X^j K_h^2(X)\sigma_+^2(X)|Z=1]$ is upper bounded by

$$\begin{aligned}\mathbb{E}[(K_h(X))^4 X^{2j}\sigma_+^4(X)|Z=1] &= \frac{1}{h^4}\int_{-\infty}^{\infty} K^4(x/h)x^{2j}\sigma_+^4(x)f_+(x)\,\mathrm{d}x \\ &= h^{2j-3}\int_{-\infty}^{\infty} K^4(u)u^{2j}\sigma_+^4(hu)f_+(hu)\,\mathrm{d}u \\ &= O(h^{2j-3}),\end{aligned}$$

where the last equality holds because $K(\cdot)$ is bounded with bounded support, $\sigma_+^2(\cdot)f_+(\cdot)$ is continuous and strictly positive at 0, and asymptotically $h \to 0$.

Thus

$$\begin{aligned}\frac{\mathrm{Var}[X^j K_h^2(X)\sigma_+^2(X)|Z=1]}{N_+} &\leqslant \frac{\mathbb{E}[X^j K_h^2(X)\sigma_+^2(X)|Z=1]}{N_+} \\ &= \frac{O(h^{2j-3})}{N_+} = O(\frac{h^{2j-2}}{hN})\cdot\frac{N}{N_+} = o_p(h^{2j-2}),\end{aligned}$$

where the last equality holds because as $N \to \infty$, $hN \to \infty$, so $h^{2j}/hN = o(h^{2j})$, and because $N/N_+ = O_p(1)$. Combining previous results,

$$\begin{aligned}G_{j,+} &= h^{j-1}\pi_j\sigma_+^2(0)f_+(0) + o(h^{j-1}) + O_p(o_p(h^{j-1})) \\ &= h^{j-1}\pi_j\sigma_+^2(0)f_+(0) + o_p(h^{j-1}).\end{aligned}$$

□

We are now ready to compute an asymptotic approximation for the bias and variance of the causal estimator $\hat{\tau}_{\mathrm{thresh}} = \hat{\tau}(0)$. Recall that as discussed in Section 3, $\hat{\tau}(0)$ can be equivalently estimated by solving two separate local linear regressions, one for the treatment group and one for the control group, rather than solving (4) and plugging the solution into (5). In this proof, we use the two separate local linear regression formulation, so that the proof more closely resembles that seen in the appendix of Imbens and Kalyanaraman (2012).

In particular, define $\mathcal{X}_+ \in \mathbb{R}^{N_+\times 2}$ and $\mathcal{X}_- \in \mathbb{R}^{N_-\times 2}$ to be the design matrices for the local linear regression restricted to the treated group and the control group respectively. That is define

$$\mathcal{X}_+ \equiv \begin{bmatrix}\mathbf{1} & (X_i)_{i\in\mathcal{Z}_+}\end{bmatrix} \quad \text{and} \quad \mathcal{X}_- \equiv \begin{bmatrix}\mathbf{1} & (X_i)_{i\in\mathcal{Z}_-}\end{bmatrix}.$$

Also define the corresponding local linear regression weight matrices by

$$W_+ \equiv \mathrm{diag}\Big(\big(K_h(X_i)\big)_{i\in\mathcal{Z}_+}\Big) \quad \text{and} \quad W_- \equiv \mathrm{diag}\Big(\big(K_h(X_i)\big)_{i\in\mathcal{Z}_-}\Big),$$

and similarly define the corresponding conditional variance matrices by

$$\Sigma_+ \equiv \mathrm{diag}\Big(\big(\sigma_+^2(X_i)\big)_{i\in\mathcal{Z}_+}\Big) \quad \text{and} \quad \Sigma_- \equiv \mathrm{diag}\Big(\big(\sigma_-^2(X_i)\big)_{i\in\mathcal{Z}_-}\Big).$$

Finally, define $\mathcal{Y}_+ \equiv (Y_i)_{i\in\mathcal{Z}_+}$, $\mathcal{Y}_- \equiv (Y_i)_{i\in\mathcal{Z}_-}$, and let $\mathbf{e}_1 = (1,0)$. The causal estimator for $\tau(0) = \mu_+(0) - \mu_-(0)$ is given by $\hat{\tau}(0) = \hat{\mu}_+(0) - \hat{\mu}_-(0)$, where $\hat{\mu}_+(0)$ and $\hat{\mu}_-(0)$ are the local linear regression estimators given by

$$\hat{\mu}_+(0) = \mathbf{e}_1^{\mathsf{T}}(\mathcal{X}_+^{\mathsf{T}}W_+\mathcal{X}_+)^{-1}\mathcal{X}_+^{\mathsf{T}}W_+\mathcal{Y}_+ \quad \text{and} \quad \hat{\mu}_-(0) = \mathbf{e}_1^{\mathsf{T}}(\mathcal{X}_-^{\mathsf{T}}W_-\mathcal{X}_-)^{-1}\mathcal{X}_-^{\mathsf{T}}W_-\mathcal{Y}_-.$$

Let $\mathcal{X} \in \mathbb{R}^{N\times 4}$ be the full design matrix whose ith row is $(1, X_i, Z_i, X_iZ_i)$, and note that the matrices $\mathcal{X}_+$ and $\mathcal{X}_-$ and the sets $\mathcal{Z}_+$ and $\mathcal{Z}_-$ are functions of the full design matrix.

In the next subsection, we compute the asymptotic approximation to the bias of the estimator $\mathbb{E}[\hat{\tau}(0)|\mathcal{X}] - \tau(0)$, and in the following subsection, we compute the asymptotic approximation to its variance $\mathrm{Var}[\hat{\tau}(0)|\mathcal{X}]$. These calculations leverage Lemmas A.1 and A.2.

### *Asymptotic approximation of the bias*

Define $B_+ \equiv \mathbb{E}[\hat{\mu}_+(0)\,|\,\mathcal{X}] - \mu_+(0)$ and $B_- \equiv \mathbb{E}[\hat{\mu}_-(0)|\mathcal{X}] - \mu_-(0)$, and note that

$$\mathbb{E}[\hat{\tau}(0)|\mathcal{X}] - \tau(0) = B_+ - B_-.$$

We will compute the asymptotic formula for $B_+$, and by an identical argument, the asymptotic formula for $B_-$ will follow.

To do this note that

$$\mathbb{E}[\hat{\mu}_+(0)|\mathcal{X}] = \mathbf{e}_1^\mathsf{T}(\mathcal{X}_+^\mathsf{T}W_+\mathcal{X}_+)^{-1}\mathcal{X}_+^\mathsf{T}W_+\mathbb{E}[\mathcal{Y}_+|\mathcal{X}] = \mathbf{e}_1^\mathsf{T}(\mathcal{X}_+^\mathsf{T}W_+\mathcal{X}_+)^{-1}\mathcal{X}_+^\mathsf{T}W_+M_+$$

where $M_+ = \big(\mu_+(X_i)\big)_{i\in\mathcal{Z}_+}$.

Since $\mu_+(\cdot)$ has at least three continuous derivatives in an open neighborhood of 0, for each $i \in \mathcal{Z}_i$,

$$\mu_+(X_i) = \mu_+(0) + \mu_+^{(1)}(0)X_i + \frac{1}{2}\mu_+^{(2)}(0)X_i^2 + T_i,$$

where $|T_i| \leqslant |X_i^3| \cdot \sup_{x\in[-|X_i|,|X_i|]} |\mu_+^{(3)}(x)|$.

So letting $T_+ = \big(T_i\big)_{i\in\mathcal{Z}_+}$ and $S_+ = \big(\mu_+^{(2)}(0)X_i^2/2\big)_{i\in\mathcal{Z}_+}$ it follows that

$$M_+ = \mathcal{X}_+\begin{bmatrix}\mu_+(0)\\ \mu_+^{(1)}(0)\end{bmatrix} + S_+ + T_+.$$

Combining previous results

$$\begin{aligned} B_+ &= e_1^\mathsf{T}(\mathcal{X}_+^\mathsf{T}W_+\mathcal{X}_+)^{-1}\mathcal{X}_+^\mathsf{T}W_+M_+ - \mu_+(0)\\ &= e_1^\mathsf{T}(\mathcal{X}_+^\mathsf{T}W_+\mathcal{X}_+)^{-1}\mathcal{X}_+^\mathsf{T}W_+(S_+ + T_+).\end{aligned}$$

Since $\nu_0\nu_2 - \nu_1^2 > 0$ (by the Cauchy-Schwartz inequality), Lemma A.1 and a first order Taylor expansion, as $h \to 0$ and $N \to \infty$, yield

$$\frac{1}{F_{0,+}F_{2,+} - F_{1,+}^2} = \frac{1}{h^2 f_+^2(0)[\nu_0\nu_2 - \nu_1^2]}\Big(\frac{1}{1+o_p(1)}\Big) = \frac{1+o_p(1)}{h^2 f_+^2(0)[\nu_0\nu_2 - \nu_1^2]}.$$

Hence, by Lemma A.1 again,

$$\begin{aligned}
\left(\frac{1}{N_+}\mathcal{X}_+^{\mathsf{T}} W_+ \mathcal{X}_+\right)^{-1} &= \begin{bmatrix} F_{0,+} & F_{1,+} \\ F_{1,+} & F_{2,+} \end{bmatrix}^{-1} \\
&= \frac{1}{F_{0,+}F_{2,+} - F_{1,+}^2} \begin{bmatrix} F_{2,+} & -F_{1,+} \\ -F_{1,+} & F_{0,+} \end{bmatrix} \\
&= \big(1 + o_p(1)\big) \begin{bmatrix} \frac{h^2[f_+(0)\nu_2 + o_p(1)]}{h^2 f_+^2(0)[\nu_0\nu_2 - \nu_1^2]} & \frac{-h[f_+(0)\nu_1 + o_p(1)]}{h^2 f_+^2(0)[\nu_0\nu_2 - \nu_1^2]} \\ \frac{-h[f_+(0)\nu_1 + o_p(1)]}{h^2 f_+^2(0)[\nu_0\nu_2 - \nu_1^2]} & \frac{f_+(0)\nu_0 + o_p(1)}{h^2 f_+^2(0)[\nu_0\nu_2 - \nu_1^2]} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\nu_2}{f_+(0)(\nu_0\nu_2 - \nu_1^2)} + o_p(1) & \frac{-\nu_1}{hf_+(0)(\nu_0\nu_2 - \nu_1^2)} + o_p(1/h) \\ \frac{-\nu_1}{hf_+(0)(\nu_0\nu_2 - \nu_1^2)} + o_p(1/h) & \frac{\nu_0}{h^2 f_+(0)(\nu_0\nu_2 - \nu_1^2)} + o_p(1/h^2) \end{bmatrix} \\
&= \begin{bmatrix} O_p(1) & O_p(1/h) \\ O_p(1/h) & O_p(1/h^2) \end{bmatrix}.
\end{aligned}$$

Now note, that since $\mu_+^{(3)}(\cdot)$ is continuous in an open neighborhood of 0, there exists an $a > 0$ for which $\sup_{x\in[-a,a]} |\mu_+^{(3)}(x)| \equiv \overline{a}_+ < \infty$. Since $K(\cdot)$ has bounded support, there exists an $\epsilon > 0$ such that for all $h \in (0, \epsilon)$, $K_h(x) = 0$ whenever $|x| > a$. Therefore for all $h < \epsilon$ and all $i$,

$$|K_h(X_i)T_i| \leqslant \left|K_h(X_i) \cdot |X_i^3| \cdot \sup_{x\in[-|X_i|,|X_i|]} |\mu_+^{(3)}(x)|\right| \leqslant \overline{a}_+ K_h(X_i)|X_i^3|.$$

Thus for all $h$ sufficiently small,

$$\frac{1}{N_+}\mathcal{X}_+^{\mathsf{T}} W_+ T_+ \leqslant \overline{a}_+ \begin{bmatrix} \frac{1}{N_+}\sum_{i\in\mathcal{Z}_+} K_h(X_i)|X_i^3| \\ F_{4,+} \end{bmatrix} = \begin{bmatrix} O_p(h^3) \\ O_p(h^4) \end{bmatrix}$$

where the bottom equality holds by Lemma A.1, and the top equality holds by the exact same argument as the proof of Lemma A.1 except with absolute values.

Combining the two previous results and multiplying through and dividing by $N_+$,

$$e_1^{\mathsf{T}}(\mathcal{X}_+^{\mathsf{T}} W_+ \mathcal{X}_+)^{-1}\mathcal{X}_+^{\mathsf{T}} W_+ T_+ = O_p(1)O_p(h^3) + O_p(1/h)O_p(h^4) = O_p(h^3) = o_p(h^2).$$

Similarly, by Lemma A.1,

$$\frac{1}{N_+}\mathcal{X}_+^{\mathsf{T}} W_+ S_+ = \frac{\mu_+^{(2)}(0)}{2}\begin{bmatrix} F_{2,+} \\ F_{3,+} \end{bmatrix} = \frac{\mu_+^{(2)}(0)}{2} f_+(0)\begin{bmatrix} \nu_2 h^2 + o_p(h^2) \\ \nu_3 h^3 + o_p(h^3) \end{bmatrix},$$

and therefore,

$$\begin{aligned}
&B_+ \\
&\quad = e_1^{\mathsf{T}}(\mathcal{X}_+^{\mathsf{T}} W_+ \mathcal{X}_+)^{-1}\mathcal{X}_+^{\mathsf{T}} W_+ (S_+ + T_+)
\end{aligned}$$

$$
\begin{aligned}
&= e_1^{\mathsf{T}}\Big(\frac{1}{N_+}\mathcal{X}_+^{\mathsf{T}}W_+\mathcal{X}_+\Big)^{-1}\Big(\frac{1}{N_+}\mathcal{X}_+^{\mathsf{T}}W_+S_+\Big) + o_p(h^2)\\
&= \frac{1}{2}\mu_+^{(2)}(0)f_+(0)\Big[\Big[\frac{\nu_2}{f_+(0)(\nu_0\nu_2-\nu_1^2)} + o_p(1)\Big][\nu_2 h^2 + o_p(h^2)]\Big]\\
&\quad + \frac{1}{2}\mu_+^{(2)}(0)f_+(0)\Big[\Big[-\frac{\nu_1}{hf_+(0)(\nu_0\nu_2-\nu_1^2)} + o_p(1/h)\Big][\nu_3 h^3 + o_p(h^3)]\Big] + o_p(h^2)\\
&= \frac{1}{2}\mu_+^{(2)}(0)\Big(\frac{\nu_2^2-\nu_3\nu_1}{\nu_0\nu_2-\nu_1^2}\Big)h^2 + o_p(h^2).
\end{aligned}
$$

A similar argument shows that

$$
B_- = \frac{1}{2}\mu_-^{(2)}(0)\Big(\frac{\nu_2^2-\nu_3\nu_1}{\nu_0\nu_2-\nu_1^2}\Big)h^2 + o_p(h^2).
$$

Hence, recalling that the bias is given by $\mathbb{E}[\hat{\tau}(0)|\mathcal{X}] - \tau(0) = B_+ - B_-$ it follows that the asymptotic formula for the bias is

$$
\text{bias}_{\hat{\tau}(0)} = \mathbb{E}[\hat{\tau}(0)|\mathcal{X}] - \tau(0) = \frac{1}{2}\Big(\mu_+^{(2)}(0) - \mu_-^{(2)}(0)\Big)\Big(\frac{\nu_2^2-\nu_3\nu_1}{\nu_0\nu_2-\nu_1^2}\Big)h^2 + o_p(h^2). \quad (36)
$$

Squaring the above formula, we recover the leading order squared-bias. As noted in Section 4, the leading order squared-bias is the first term in (12).

### *Asymptotic approximation of the variance*

Defining $V_\pm = \text{Var}(\hat{\mu}_\pm(0)\,|\,\mathcal{X})$, the variance for our causal estimator is

$$
\text{Var}[\hat{\tau}(0)\,|\,\mathcal{X}] = V_+ + V_- + 2\text{Cov}[\hat{\mu}_+(0), \hat{\mu}_-(0)\,|\,\mathcal{X}] = V_+ + V_-,
$$

where the last equality holds because $(X_i, Y_{i+}, Y_{i-})$ are IID by assumption making $\mathcal{Y}_+$ and $\mathcal{Y}_-$ independent conditionally on the treatment assignments which then implies that $\hat{\mu}_\pm(0)$ are independent conditionally on $\mathcal{X}$.

We now compute an asymptotic formula for $V_+$ and by an identical argument, the asymptotic formula for $V_-$ will follow.

First note that

$$
\begin{aligned}
V_+ &= \text{Var}[\mathbf{e}_1^{\mathsf{T}}(\mathcal{X}_+^{\mathsf{T}}W_+\mathcal{X}_+)^{-1}\mathcal{X}_+^{\mathsf{T}}W_+\mathcal{Y}_+|\mathcal{X}]\\
&= \mathbf{e}_1^{\mathsf{T}}(\mathcal{X}_+^{\mathsf{T}}W_+\mathcal{X}_+)^{-1}\mathcal{X}_+^{\mathsf{T}}W_+\text{Var}[\mathcal{Y}_+|\mathcal{X}]W_+^{\mathsf{T}}\mathcal{X}_+(\mathcal{X}_+^{\mathsf{T}}W_+\mathcal{X}_+)^{-1}\mathbf{e}_1\\
&= \mathbf{e}_1^{\mathsf{T}}(\mathcal{X}_+^{\mathsf{T}}W_+\mathcal{X}_+)^{-1}\mathcal{X}_+^{\mathsf{T}}W_+\Sigma_+W_+\mathcal{X}_+(\mathcal{X}_+^{\mathsf{T}}W_+\mathcal{X}_+)^{-1}\mathbf{e}_1
\end{aligned}
$$

Now the middle factor rescaled by $1/N_+$ is

$$
\frac{1}{N_+}\mathcal{X}_+^{\mathsf{T}}W_+\Sigma_+W_+\mathcal{X}_+ = \begin{bmatrix} G_{0,+} & G_{1,+} \\ G_{1,+} & G_{2,+}\end{bmatrix}
$$

and recall from the previous subsection that

$$
\Big(\frac{1}{N_+}\mathcal{X}_+^{\mathsf{T}}W_+\mathcal{X}_+\Big)^{-1} = \frac{1}{F_{0,+}F_{2,+}-F_{1,+}^2}\begin{bmatrix} F_{2,+} & -F_{1,+} \\ -F_{1,+} & F_{0,+}\end{bmatrix}.
$$

Thus,

$$
\begin{aligned}
&N_+V_+ \\
&= \frac{1}{(F_{0,+}F_{2,+} - F_{1,+}^2)^2}\mathbf{e}_1^\mathsf{T}\begin{bmatrix} F_{2,+} & -F_{1,+} \\ -F_{1,+} & F_{0,+} \end{bmatrix}\begin{bmatrix} G_{0,+} & G_{1,+} \\ G_{1,+} & G_{2,+} \end{bmatrix}\begin{bmatrix} F_{2,+} & -F_{1,+} \\ -F_{1,+} & F_{0,+} \end{bmatrix}\mathbf{e}_1 \\
&= \frac{F_{2,+}^2 G_{0,+} - 2F_{1,+}F_{2,+}G_{1,+} + F_{1,+}^2 G_{2,+}}{(F_{0,+}F_{2,+} - F_{1,+}^2)^2} \\
&= \frac{\sigma_+^2(0)}{f_+(0)} \cdot \frac{\left(\nu_2^2\pi_0 h^3 - 2\nu_1\nu_2\pi_1 h^3 + \nu_1^2\pi_2 h^3\right)[1 + o_p(1)]^3}{(\nu_0\nu_2 h^2 - \nu_1^2 h^2)^2[1 + o_p(1)]^4} \\
&= \frac{\sigma_+^2(0)}{hf_+(0)}\Big(\frac{\nu_2^2\pi_0 - 2\nu_1\nu_2\pi_1 + \nu_1^2\pi_2}{(\nu_0\nu_2 - \nu_1^2)^2} + o_p(1)\Big) \\
&= \frac{2\sigma_+^2(0)\Pr(Z_i = 1)}{hf(0)}\Big(\frac{\nu_2^2\pi_0 - 2\nu_1\nu_2\pi_1 + \nu_1^2\pi_2}{(\nu_0\nu_2 - \nu_1^2)^2} + o_p(1)\Big).
\end{aligned}
$$

Thus dividing through by $N_+$ and rearranging terms,

$$V_+ = \frac{2\sigma_+^2(0)}{Nhf(0)}\Big(\frac{\nu_2^2\pi_0 - 2\nu_1\nu_2\pi_1 + \nu_1^2\pi_2}{(\nu_0\nu_2 - \nu_1^2)^2} + o_p(1)\Big) \cdot \frac{N\Pr(Z_i = 1)}{N_+}.$$

By the weak law of large numbers $N\Pr(Z_i = 1)/N_+ = (1 + o_p(1))$ and therefore,

$$V_+ = \frac{2\sigma_+^2(0)}{Nhf(0)}\Big(\frac{\nu_2^2\pi_0 - 2\nu_1\nu_2\pi_1 + \nu_1^2\pi_2}{(\nu_0\nu_2 - \nu_1^2)^2}\Big) + o_p\Big(\frac{1}{Nh}\Big).$$

A similar argument shows that

$$V_- = \frac{2\sigma_-^2(0)}{Nhf(0)}\Big(\frac{\nu_2^2\pi_0 - 2\nu_1\nu_2\pi_1 + \nu_1^2\pi_2}{(\nu_0\nu_2 - \nu_1^2)^2}\Big) + o_p\Big(\frac{1}{Nh}\Big),$$

and hence because $\mathrm{Var}[\hat{\tau}(0)|\mathcal{X}] = V_+ + V_-$,

$$\mathrm{Var}[\hat{\tau}(0)|\mathcal{X}] = \frac{\sigma_+^2(0) + \sigma_-^2(0)}{Nhf(0)}\Big(\frac{2\nu_2^2\pi_0 - 4\nu_1\nu_2\pi_1 + 2\nu_1^2\pi_2}{(\nu_0\nu_2 - \nu_1^2)^2}\Big) + o_p\Big(\frac{1}{Nh}\Big). \quad (37)$$

As remarked in Section 4, the leading order variance matches the second term in formula (12).

### *Asymptotic expression for mean squared error*

To complete the proof of the asymptotic expression for the tie-breaker design MSE in Lemma 2, note that

$$\mathrm{MSE}_{\mathrm{TBD}}(h, N)$$

$$
\begin{aligned}
&= \mathbb{E}\big[\big(\hat{\tau}(0) - \tau(0)\big)^2 \big| \mathcal{X}\big] \\
&= \text{bias}^2_{\hat{\tau}(0)} + \text{Var}[\hat{\tau}(0)|\mathcal{X}] \\
&= C_1\big(\mu_+^{(2)}(0) - \mu_-^{(2)}(0)\big)^2 h^4 + o_p(h^4) + \frac{C_2}{Nh}\Big(\frac{\sigma_+^2(0) + \sigma_-^2(0)}{f(0)}\Big) + o_p\Big(\frac{1}{Nh}\Big) \\
&= \text{AMSE}_{\text{TBD}}(h, N) + o_p\Big(h^4 + \frac{1}{Nh}\Big)
\end{aligned}
$$

where the second last equality holds from combining equations (37), (36), and (11), while the last equality holds from definition (12).

***Asymptotically optimal bandwidth expression***

To complete the proof of Lemma 2, recall that we define

$$h_{\text{opt,TBD}}(N) = \arg\min_h \text{AMSE}_{\text{TBD}}(h, N)$$

and by (12),

$$\text{AMSE}_{\text{TBD}}(h, N) \equiv C_1\big(\mu_+^{(2)}(0) - \mu_-^{(2)}(0)\big)^2 h^4 + \frac{C_2}{Nh}\Big(\frac{\sigma_+^2(0) + \sigma_-^2(0)}{f(0)}\Big).$$

Because by assumption $\mu_+^{(2)}(0) \neq \mu_-^{(2)}(0)$ and $f(0) > 0$, a simple calculus exercise shows that

$$h_{\text{opt,TBD}}(N) = \Big(\frac{C_2}{4C_1}\Big)^{1/5}\Big(\frac{\sigma_+^2(0) + \sigma_-^2(0)}{f(0)\big(\mu_+^{(2)}(0) - \mu_-^{(2)}(0)\big)^2}\Big)^{1/5} N^{-1/5}. \quad \square$$

## Appendix B: Proof of Theorem 1

First note that by plugging (17) into (7), $\text{AMSE}_{\text{RDD}}\big(\hat{h}_{\text{opt,RDD}}(N), N\big)$ equals

$$
\begin{aligned}
&\tilde{C}_1\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2 \hat{h}^4_{\text{opt,RDD}}(N) + \frac{\tilde{C}_2}{N\hat{h}_{\text{opt,RDD}}(N)}\Big(\frac{\sigma_+^2(t) + \sigma_-^2(t)}{f(t)}\Big) \\
&= \big(4\tilde{C}_1\tilde{C}_2^4\big)^{1/5} N^{-4/5}\Big[\frac{1}{4}\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2 \hat{\gamma}^4_{\text{RDD}} + \frac{\sigma_+^2(t) + \sigma_-^2(t)}{\hat{\gamma}_{\text{RDD}} f(t)}\Big] \\
&= \big(4\tilde{C}_1\tilde{C}_2^4\big)^{1/5} N^{-4/5}\Big[\frac{1}{4}\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2 \big(\gamma + o_p(1)\big)^4 + \frac{\sigma_+^2(t) + \sigma_-^2(t)}{\big(\gamma + o_p(1)\big) f(t)}\Big] \\
&= \big(4\tilde{C}_1\tilde{C}_2^4\big)^{1/5} N^{-4/5}\Big[\frac{1}{4}\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2 \gamma^4 + \frac{\sigma_+^2(t) + \sigma_-^2(t)}{\gamma f(t)} + o_p(1)\Big] \\
&= \big(4\tilde{C}_1\tilde{C}_2^4\big)^{1/5} N^{-4/5}\big[\alpha + o_p(1)\big],
\end{aligned}
$$

where the third equality holds since $\hat{\gamma}_{\text{RDD}} \xrightarrow{p} \gamma$ and the last inequality uses definitions (15) and (16).

Plugging (18) into (12), an identical argument that uses $\hat{\gamma}_{\text{TBD}} \xrightarrow{p} \gamma$ and definitions (15) and (16) yields

$$\text{AMSE}_{\text{TBD}}\big(\hat{h}_{\text{opt,TBD}}(\theta N), \theta N\big) = \big(4C_1C_2^4\big)^{1/5}(\theta N)^{-4/5}\big[\alpha + o_p(1)\big].$$

In applying Lemma 1, it will be helpful to note that letting $c = \big(\tilde{C}_2/(4\tilde{C}_1)\big)^{1/5}$,

$$\begin{aligned}
\hat{h}^4_{\text{opt,RDD}}(N) + \frac{1}{N\hat{h}_{\text{opt,RDD}}(N)} &= c^4\hat{\gamma}^4_{\text{RDD}}N^{-4/5} + \frac{1}{c\hat{\gamma}_{\text{RDD}}}N^{-4/5} \\
&= \Big(c^4\big(\gamma + o_p(1)\big)^4 + \frac{1}{c\big(\gamma + o_p(1)\big)}\Big)N^{-4/5} \\
&= \big(c^4\gamma + \frac{1}{c\gamma} + o_p(1)\big)N^{-4/5} \\
&= O_p(N^{-4/5}).
\end{aligned}$$

Therefore under the conditions of Theorem 1, by applying Lemmas 1 and 2 and combining the previous results,

$$\begin{aligned}
&\frac{\text{MSE}_{\text{RDD}}\big(\hat{h}_{\text{opt,RDD}}(N), N\big)}{\text{MSE}_{\text{TBD}}\big(\hat{h}_{\text{opt,TBD}}(\theta N), \theta N\big)} \\
&\quad= \frac{\text{AMSE}_{\text{RDD}}\big(\hat{h}_{\text{opt,RDD}}(N), N\big) + o_p(O_p(N^{-4/5}))}{\text{AMSE}_{\text{TBD}}\big(\hat{h}_{\text{opt,TBD}}(\theta N), \theta N\big) + o_p(O_p(N^{-4/5}))} \\
&\quad= \frac{\big(4\tilde{C}_1\tilde{C}_2^4\big)^{1/5}N^{-4/5}\big[\alpha + o_p(1)\big] + N^{-4/5}o_p(1)}{\big(4C_1C_2^4\big)^{1/5}(\theta N)^{-4/5}\big[\alpha + o_p(1)\big] + N^{-4/5}o_p(1)} \\
&\quad= \frac{\big(4\tilde{C}_1\tilde{C}_2^4\big)^{1/5}\big[\alpha + o_p(1)\big] + o_p(1)}{\big(4C_1C_2^4\big)^{1/5}\theta^{-4/5}\big[\alpha + o_p(1)\big] + o_p(1)} \\
&\quad= \theta^{4/5}\Big(\frac{\tilde{C}_1\tilde{C}_2^4}{C_1C_2^4}\Big)^{1/5} + o_p(1). \quad \square
\end{aligned}$$

## Appendix C: Extensions to assignment probability $p \neq 1/2$

In this appendix we consider the MSE of a generalization of the 3-level TBD given by (1), where we allow the assignment probabilities to $Z = 1$ and $Z = -1$ to differ from $1/2$ within the interval of experimentation. In particular, we consider a design with the following assignment probabilities

$$\Pr(Z_i = 1 \,|\, x_i) = \begin{cases} 0, & x_i - t \leqslant -\Delta \\ p, & |x_i - t| < \Delta \\ 1, & x_i - t \geqslant \Delta, \end{cases} \tag{38}$$

for some $p \in (0,1)$. Under this more general version of the TBD, one can show that the AMSE formula given in (12) changes to

$$\mathrm{AMSE}_{\mathrm{TBD}(p)}(h,N) \equiv C_1\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2 h^4 + \frac{C_2}{Nh}\Big(\frac{\frac{1}{2p}\sigma_+^2(t) + \frac{1}{2(1-p)}\sigma_-^2(t)}{f(t)}\Big), \tag{39}$$

and that the following lemma holds.

**Lemma C.1.** *Under the conditions of Lemma 2, with the exception that the assignment probabilities follow (38) rather than (1), the mean squared error in estimating $\tau_{\mathrm{thresh}}$ is given by*

$$\mathrm{MSE}_{\mathrm{TBD}(p)}(h,N) = \mathrm{AMSE}_{\mathrm{TBD}(p)}(h,N) + o_p\Big(h^4 + \frac{1}{Nh}\Big),$$

*and the asymptotically optimal bandwidth, defined by* $\arg\min_h \mathrm{AMSE}_{\mathrm{TBD}(p)}(h,N)$ *is*

$$h_{\mathrm{opt,TBD}(p)}(N) = \Big(\frac{C_2}{4C_1}\Big)^{1/5}\Big(\frac{\frac{1}{2p}\sigma_+^2(t) + \frac{1}{2(1-p)}\sigma_-^2(t)}{f(t)\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2}\Big)^{1/5} N^{-1/5}. \tag{40}$$

*Proof.* The proof is identical to the proof of Lemma 2 in Appendix A, except that the formulas for the functions $f_\pm(\cdot)$ presented at (34) are instead given by

$$f_+(x) = \begin{cases} \frac{f(x)}{\Pr(Z_i=1)}, & x \geqslant \Delta \\ \frac{pf(x)}{\Pr(Z_i=1)}, & |x| < \Delta \\ 0, & x \leqslant -\Delta \end{cases} \quad \text{and} \quad f_-(x) = \begin{cases} 0, & x \geqslant \Delta \\ \frac{(1-p)f(x)}{\Pr(Z_i=-1)}, & |x| < \Delta \\ \frac{f(x)}{\Pr(Z_i=-1)}, & x \leqslant -\Delta. \end{cases}$$

The changes in the formulas for $f_\pm(\cdot)$ do not affect the leading order bias formula but do affect the leading order variance and optimal bandwidth formulas. □

Analogously to the empirical bandwidth choice given by (18), an investigator running a TBD of the form (38) seeking mean squared optimal estimation of $\tau_{\mathrm{thresh}}$ would ultimately use the bandwidth

$$\hat{h}_{\mathrm{opt,TBD}(p)}(N) = \Big(\frac{C_2}{4C_1}\Big)^{1/5}\hat{\gamma}_{\mathrm{TBD}(p)} N^{-1/5},$$

where $\hat{\gamma}_{\mathrm{TBD}(p)}$ is some consistent estimator of the quantity

$$\Big(\frac{\frac{1}{2p}\sigma_+^2(t) + \frac{1}{2(1-p)}\sigma_-^2(t)}{f(t)\big(\mu_+^{(2)}(t) - \mu_-^{(2)}(t)\big)^2}\Big)^{1/5}.$$

To derive an analogue of Theorem 1 for a TBD of the form (38), it is convenient to define the relative variance measure

$$r(t) \equiv \frac{\sigma_+^2(t)}{\sigma_+^2(t) + \sigma_-^2(t)}. \tag{41}$$

The following theorem compares the RDD with $N$ points to a TBD of the form (38) with $\theta N$ points for some $\theta > 0$.

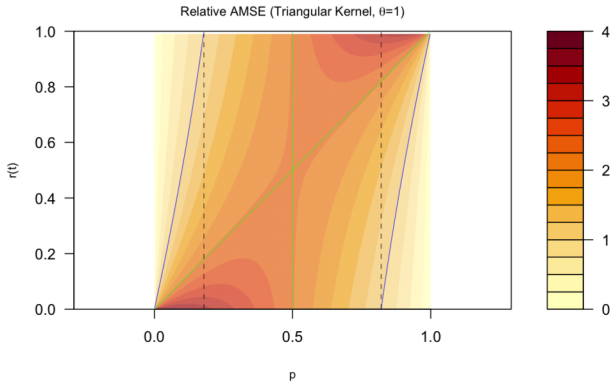



**Fig C.1:** Relative AMSE given by Theorem C.1 as function of $p$ and $r(t)$. Here we consider the same sample size for each design ($\theta = 1$), and that the most favorable kernel for the RDD (the triangular kernel) is to be used. The color scale for the contour plot is displayed on the right, with the added blue lines denoting a level of 1. Outside the blue lines, the RDD performs better than the TBD asymptotically. The vertical black lines, between which the relative AMSE is always greater than 1, are placed at $p = \theta_*/2$ and $p = 1 - \theta_*/2$, for $\theta_* = 60.46618^{-1/4}$. The pair of green lines give the boundaries of the two triangular regions in which the TBD with assignment probabilities (38) has a smaller AMSE than the default TBD with $p = 1/2$ has.

**Theorem C.1.** *Under the assumptions of Theorem 1, as* $N \to \infty$

$$\frac{\mathrm{MSE}_{\mathrm{RDD}}\big(\hat{h}_{\mathrm{opt,RDD}}(N), N\big)}{\mathrm{MSE}_{\mathrm{TBD}(p)}\big(\hat{h}_{\mathrm{opt,TBD}(p)}(\theta N), \theta N\big)} \xrightarrow{p} \theta^{4/5}\Big(\frac{\tilde{C}_1\tilde{C}_2^4}{C_1C_2^4}\Big)^{1/5}\Big(\frac{r(t)}{2p} + \frac{1-r(t)}{2(1-p)}\Big)^{-4/5} \tag{42}$$

*holds for any tie-breaker design of the form* (38) *with* $\Delta > 0$.

*Proof.* The proof is the same as that of Theorem 1, except the formulas for $\mathrm{AMSE}_{\mathrm{TBD}(p)}(h, N)$, $h_{\mathrm{opt,TBD}(p)}(N)$, and $\hat{h}_{\mathrm{opt,TBD}(p)}(N)$ are different in the setting of a tie-breaker design of the form (38). □

We visualize the implications of Theorem C.1 in Figure C.1. While the figure focuses on the case of the triangular kernel with $\theta = 1$, the relative AMSE as a function of $p$ and $r(t)$ will be a scalar multiple of that shown in Figure C.1 for different kernels and values of $\theta$. We also explicitly list a few notable implications of Theorem C.1 below.

1. The AMSE for the TBD is minimized by $p = \sigma_+(t)/\big(\sigma_-(t) + \sigma_+(t)\big)$.
2. If either $r(t) \leqslant p \leqslant 1/2$ or $1/2 \leqslant p \leqslant r(t)$, then

$$\Big(\frac{r(t)}{2p} + \frac{1-r(t)}{2(1-p)}\Big)^{-4/5} \geqslant 1.$$

As a result, whenever $p$ is between $r(t)$ and $1/2$ inclusive, the advantage of a TBD over an RDD (in terms of relative asymptotic MSE in estimation of

$\tau_{\text{thresh}}$) is at least as large as that described in Table 1, which corresponds to the $p = 1/2$ case.

3. Suppose that a TBD and RDD are to be conducted with the same sample size $N$ (i.e. $\theta = 1$), and that under both designs the asymptotically MSE optimal bandwidth is to be used. Then, regardless of the unknown value of $r(t) \in [0, 1]$, for any choice of $p \in (\theta_*/2, 1 - \theta_*/2)$ with $\theta_*$ defined at (20), the TBD given by (38) will have lower AMSE than the RDD. Conversely, if $p \notin [\theta_*/2, 1-\theta_*/2]$, there will exist a range of values of $r(t)$ contained in $[0, 1]$ for which the TBD will have higher AMSE than the RDD. Note that amongst the common kernel choices considered in Table 1, $\theta_*$ is largest for the triangular kernel with a value of precisely $60.46618^{-1/4}$, suggesting that for any of these kernel choices the TBD will have lower AMSE than an RDD in estimating $\tau_{\text{thresh}}$ as long as $p \in [0.18, 0.82]$, but outside that range, the TBD may have larger AMSE depending on the kernel choice and the unknown value of $r(t)$.

The first fact may be difficult to leverage in practice because at the time an experiment is being designed, the investigator is unlikely to have outcome data from both treatment and control units that they can use to estimate the ratio $\sigma_+(t)/\sigma_-(t)$. If the investigator has some prior knowledge about whether or not $\sigma_+(t)/\sigma_-(t)$ is smaller or greater than 1, they could potentially leverage the second fact to design a TBD that has lower AMSE than a TBD with the default of $p = 1/2$ has. The third statement is true for any value of $r(t)$ or $\sigma_+(t)/\sigma_-(t)$, suggesting that a design with $p \notin [0.18, 0.82]$ should only be considered with caution.

## Appendix D: Bandwidth that removes leading order bias

In the asymptotic analysis from Section 4, the bias for the local linear estimator of $\hat{\tau}_{\text{thresh}}$ was $O_p(h^2)$. However, that analysis relied on a simplification that arose by assuming that $h \to 0$ as $N \to \infty$. In this appendix, we derive an asymptotic expression for the bias of $\hat{\tau}_{\text{thresh}}$ that does not rely on the assumption that $h \to 0$ as $N \to \infty$. In this expression for the bias, the leading order term is quadratic in $h$. We also show that for some value of $h/\Delta$, which depends only on the kernel $K(\cdot)$, the quadratic bias term vanishes. In Table D.1, we present a numerical approximation of the value of $h/\Delta$ that removes the leading order bias for common kernel choices.

We consider the following regularity conditions, which were not needed in the main text.

(I) The mean functions $\mu_\pm(\cdot)$ are three times differentiable. Further, the third derivatives of the mean functions, $\mu_\pm^{(3)}(\cdot)$, are bounded.
(II) The density $f$ of the assignment variable satisfies $|f(x) - f(t)| \leqslant L|x - t|$ for some $L < \infty$.

In the proof of Lemma 2, where we computed the leading order bias, we did not need to assume (I) and (II) because we supposed that $h \to 0$ as $N \to \infty$. In

the following theorem about the leading order bias, because we do not assume $h \to 0$, we assume (I) and (II) in addition to (i)–(vi).

**Theorem D.1.** *Suppose conditions (i)–(vi) from the main text and conditions (I) and (II) hold, and that $N \to \infty$. Let $\Delta > 0$ be fixed. Under the tie-breaker design defined by (1) and estimation of $\tau_{\text{thresh}}$ according to (4) and (5) for some bandwidth $h > 0$, the bias of $\hat{\tau}_{\text{thresh}}$ is given by*

$$\frac{\mu_+^{(2)}(t) - \mu_-^{(2)}(t)}{2}\Big(\frac{\nu_2^2 - 4\int_{\Delta/h}^{\infty} uK(u)\,\mathrm{d}u \int_{\Delta/h}^{\infty} u^3 K(u)\,\mathrm{d}u}{\nu_0\nu_2 - 4[\int_{\Delta/h}^{\infty} uK(u)\,\mathrm{d}u]^2}\Big)h^2 + O(h^3) + O_p\Big(\frac{1}{\sqrt{N}}\Big), \tag{43}$$

*where $\nu_j$ is defined at (10). As a result, the leading order bias term is equal to 0 whenever the bandwidth $h$ is chosen so that*

$$\int_{\Delta/h}^{\infty} uK(u)\,\mathrm{d}u \int_{\Delta/h}^{\infty} u^3 K(u)\,\mathrm{d}u = \Big(\int_0^{\infty} u^2 K(u)\,\mathrm{d}u\Big)^2. \tag{44}$$

*Moreover, there must exist an $h > \Delta > 0$ that satisfies (44).*

*Proof.* Fix $\Delta > 0$ and suppose without loss of generality that $t = 0$. Define $K_h(\cdot)$, $\mathcal{Z}_\pm$, $N_\pm$, $F_{j,\pm}$ for $j = 0, 1, \ldots, 4$, $f_\pm(\cdot)$, $\mathcal{X}_\pm$, $W_\pm$, $\mathcal{Y}_\pm$, $\hat{\mu}_\pm(0)$, and $B_\pm$, according to the same definitions presented in Appendix A.

We will first derive a formula for $F_{j,\pm}$ that is similar to that given in Lemma A.1, except here we will not rely on a simplification that occurs in the asymptotic regime where $h \to 0$, eventually dropping below $\Delta$.

For any nonnegative integer $j$, because the samples are IID, we get

$$\begin{aligned}
F_{j,+} &= \frac{1}{N_+}\sum_{i\in\mathcal{Z}_+} X_i^j K_h(X_i) \\
&= \mathbb{E}[X^j K_h(X) | Z = 1] + O_p\Big(\sqrt{\mathrm{Var}[X^j K_h(X)|Z=1]/N_+}\Big) \\
&= \frac{1}{h}\int_{-\infty}^{\infty} x^j K(x/h) f_+(h)\,\mathrm{d}x + O_p\big(1/\sqrt{N_+}\big) \\
&= h^j \int_{-\infty}^{\infty} u^j K(u) f_+(hu)\,\mathrm{d}u + O_p(1/\sqrt{N}) \\
&= \frac{h^j\big(\int_{-\Delta/h}^{\Delta/h} u^j K(u) f(hu)\,\mathrm{d}u + 2\int_{\Delta/h}^{\infty} u^j K(u) f(hu)\,\mathrm{d}u\big)}{2\Pr(Z_i = 1)} + O_p(1/\sqrt{N}).
\end{aligned}$$

To simplify this expression further, observe that, for any $a, b \in [-\infty, \infty]$ and

nonnegative integer $j$,

$$\begin{aligned}\left|\int_a^b u^j K(u)\big[f(hu)-f(0)\big]\,\mathrm{d}u\right| &= \left|h\int_a^b u^{j+1}K(u)\Big[\frac{f(hu)-f(0)}{hu}\Big]\,\mathrm{d}u\right| \\ &\leqslant h\int_a^b \left|\frac{f(hu)-f(0)}{hu}\right|\cdot|u^{j+1}|K(u)\,\mathrm{d}u \\ &\leqslant Lh\int_a^b |u^{j+1}|K(u)\,\mathrm{d}u \\ &= O(h),\end{aligned}$$

where above the last two steps hold by Assumptions (II) and (iv) respectively, and as a consequence, for any $a<b$,

$$\begin{aligned}\int_a^b u^j K(u)f(hu)\,\mathrm{d}u &= \int_a^b u^j K(u)f(0)\,\mathrm{d}u + \int_a^b u^j K(u)\big[f(hu)-f(0)\big]\,\mathrm{d}u \\ &= f(0)\int_a^b u^j K(u)\,\mathrm{d}u + O(h).\end{aligned}$$

Combining this with the above formula for $F_{j,+}$,

$$\begin{aligned}F_{j,+} &= \frac{h^j f(0)\big(\int_{-\Delta/h}^{\Delta/h} u^j K(u)\,\mathrm{d}u + 2\int_{\Delta/h}^{\infty} u^j K(u)\,\mathrm{d}u + O(h)\big)}{2\Pr(Z_i=1)} + O_p(1/\sqrt{N}) \\ &= h^j f_+(0)\nu_{j,+}(\Delta/h) + O(h^{j+1}) + O_p(1/\sqrt{N}),\end{aligned}$$

where

$$\nu_{j,+}(\delta) \equiv \int_{-\delta}^{\delta} u^j K(u)\,\mathrm{d}u + 2\int_{\delta}^{\infty} u^j K(u)\,\mathrm{d}u.$$

If we define $\nu_{j,-}(\delta) \equiv \int_{-\delta}^{\delta} u^j K(u)\,\mathrm{d}u + 2\int_{-\infty}^{-\delta} u^j K(u)\,\mathrm{d}u$, a similar argument yields an analogous formula for $F_{j,-}$, and hence

$$F_{j,\pm} = h^j f_\pm(0)\nu_{j,\pm}(\Delta/h) + O(h^{j+1}) + O_p(1/\sqrt{N}) \quad \text{for all } j\in\mathbb{Z}_{\geqslant 0}. \tag{45}$$

By noting the similarities between the formula for $F_{j,+}$ in Lemma A.1 and that from (45), the formula for $B_+$ can be derived by the same argument presented in Appendix A, by simply replacing $\nu_j$ terms with $\nu_{j,+}(\Delta/h)$, $o_p(h^j)$ terms with $O(h^{j+1})+O_p(1/\sqrt{N})$ terms, and intermediary $O_p(h^j)$ terms with $O(h^j)+O_p(1/\sqrt{N})$, for any integer $j$. Because we now no longer assume $h\to 0$ as $N\to\infty$, there are three steps where we have to use a slightly different argument from that presented in Appendix A. First, since we no longer assume that $h\to 0$, the 2nd order Taylor expansion of the mean functions with the remainder terms $T_i$ require that $\mu_\pm^{(2)}(\cdot)$ is continuous everywhere, rather than merely in a neighborhood of $t$ which is given by Assumption (iii). Assumption (I) that $\mu_\pm^{(3)}(\cdot)$ is bounded, guarantees that $\mu_\pm^{(2)}(\cdot)$ is continuous everywhere because differentiability implies continuity. Second, we can use Assumption (I) that $\mu_\pm^{(3)}(\cdot)$ is bounded to show that there exists an $\bar{a}_+<\infty$ for which

$K_h(X_i)\sup_{x\in[-X_i,X_i]}|\mu_+^{(3)}(x)| \leqslant \bar{a}_+ K_h(X_i)$ for any $h>0$. In Appendix A, we did not need Assumption (I) since we supposed that $h\to 0$, so it was enough to show that inequality holds for any sufficiently small $h$. Third, the argument involves a first order Taylor expansion of the quantity $(F_{0,+}F_{2,+}-F_{1,+}^2)^{-1}$. In the current setting where we do not suppose $h\to 0$ as $N\to\infty$, we must now ascertain that with probability approaching 1 as $N\to\infty$, $F_{0,+}F_{2,+}-F_{1,+}^2$ is positive and bounded away from zero. Recalling that up to an $O_p(1/\sqrt{N})$ term, $F_{j,+}$, equals $h^j\int_{-\infty}^{\infty}u^jK(u)f_+(hu)\,\mathrm{d}u$, by the Cauchy-Schwartz inequality it follows that

$$\begin{aligned}
F_{1,+}^2 &= h^2\bigg(\int_{-\infty}^{\infty}uK(u)f_+(hu)\,\mathrm{d}u\bigg)^2+O_p(1/\sqrt{N})\\
&= h^2\bigg(\int_{-\infty}^{\infty}\sqrt{K(u)f_+(hu)}\cdot\sqrt{u^2K(u)f_+(hu)}\,\mathrm{d}u\bigg)^2+O_p(1/\sqrt{N})\\
&< h^2\int_{-\infty}^{\infty}K(u)f_+(hu)\,\mathrm{d}u\int_{-\infty}^{\infty}u^2K(u)f_+(hu)\,\mathrm{d}u+O_p(1/\sqrt{N})\\
&= F_{0,+}F_{2,+}+O_p(1/\sqrt{N}).
\end{aligned}$$

Hence with probability converging to 1 as $N\to\infty$, $F_{0,+}F_{2,+}-F_{1,+}^2$ is positive and bounded away from zero. Consequently, the following first order Taylor expansion holds for all $h>0$

$$\begin{aligned}
\frac{h^2f_+^2(0)}{F_{0,+}F_{2,+}-F_{1,+}^2} &= \frac{1}{[\nu_{0,+}(\Delta/h)\nu_{2,+}(\Delta/h)-\nu_{1,+}(\Delta/h)^2]+O(h)+O_p(1/\sqrt{N})}\\
&= \frac{1}{\nu_{0,+}(\Delta/h)\nu_{2,+}(\Delta/h)-\nu_{1,+}(\Delta/h)^2}+O(h)+O_p(1/\sqrt{N}),
\end{aligned}$$

where the denominator in the first term is positive and bounded away from 0 by a similar Cauchy-Schwartz argument that leverages symmetry of $K$.

By the same derivation of the bias formula presented in Appendix A, with the exceptions in the argument noted above, the bias of $\hat{\mu}_+(0)$ is given by

$$B_+=\frac{1}{2}\mu_+^{(2)}(0)\Big(\frac{[\nu_{2,+}(\Delta/h)]^2-\nu_{1,+}(\Delta/h)\nu_{3,+}(\Delta/h)}{\nu_{0,+}(\Delta/h)\nu_{2,+}(\Delta/h)-[\nu_{1,+}(\Delta/h)]^2}\Big)h^2+O(h^3)+O_p(1/\sqrt{N}).$$

A similar argument shows that,

$$B_-=\frac{1}{2}\mu_-^{(2)}(0)\Big(\frac{[\nu_{2,-}(\Delta/h)]^2-\nu_{1,-}(\Delta/h)\nu_{3,-}(\Delta/h)}{\nu_{0,-}(\Delta/h)\nu_{2,-}(\Delta/h)-[\nu_{1,-}(\Delta/h)]^2}\Big)h^2+O(h^3)+O_p(1/\sqrt{N}).$$

Now note that the bias of $\hat{\tau}_{\mathrm{thresh}}$ is given by $B_+-B_-$. To simplify the expression for $B_+-B_-$, observe that by symmetry of $K(\cdot)$, for any $h>0$

$$\nu_{j,-}(\Delta/h)=\begin{cases}\nu_{j,+}(\Delta/h)=\nu_j & \text{for even } j,\\ -\nu_{j,+}(\Delta/h)=-2\int_{\Delta/h}^{\infty}u^jK(u)\,\mathrm{d}u & \text{for odd } j.\end{cases}$$

The $\nu_j$ above is defined at (10). Hence, subtracting $B_-$ from $B_+$ the bias for $\hat{\tau}_{\text{thresh}}$ is given by

$$\frac{h^2}{2}\big(\mu_+^{(2)}(0)-\mu_-^{(2)}(0)\big)\frac{\nu_2^2-4\int_{\Delta/h}^{\infty}uK(u)\,\mathrm{d}u\int_{\Delta/h}^{\infty}u^3K(u)\,\mathrm{d}u}{\nu_0\nu_2-4[\int_{\Delta/h}^{\infty}uK(u)\,\mathrm{d}u]^2}+O(h^3)+O_p(1/\sqrt{N}).$$

Since we supposed $t=0$ without loss of generality, this proves (43).

Now, since $\nu_0\nu_2-4[\int_{\Delta/h}^{\infty}uK(u)\,\mathrm{d}u]^2$ is bounded away from zero, the leading order term in $h$ in formula (43) for the bias of $\hat{\tau}_{\text{thresh}}$ is zero whenever $h$ satisfies (44).

To prove the final claim, that there exists an $h>\Delta$ solving (44), define

$$\varrho(h)\equiv 4\int_{\Delta/h}^{\infty}uK(u)\,\mathrm{d}u\int_{\Delta/h}^{\infty}u^3K(u)\,\mathrm{d}u \quad \text{for } h>0.$$

Clearly by boundedness and continuity of $K(\cdot)$, $\varrho(\cdot)$ is continuous on $[0,\infty)$. Recalling Assumption (iv) that $K(\cdot)$ is supported on $[-1,1]$, $\varrho(\Delta)=0$. Since $\nu_2>0$, $\varrho(\Delta)<\nu_2^2$. Also note that by symmetry of the kernel and the Cauchy-Schwartz inequality,

$$\begin{aligned}\nu_2^2&=\left(2\int_0^{\infty}u^2K(u)\,\mathrm{d}u\right)^2\\&=4\left(\int_0^{\infty}\sqrt{uK(u)}\cdot\sqrt{u^3K(u)}\,\mathrm{d}u\right)^2\\&<4\int_0^{\infty}uK(u)\,\mathrm{d}u\int_0^{\infty}u^3K(u)\,\mathrm{d}u\\&=\lim_{h\to\infty}\varrho(h).\end{aligned}$$

Since $\varrho(\Delta)<\nu_2^2<\lim_{h\to\infty}\varrho(h)$, and since $\varrho(\cdot)$ is continuous on $[\Delta,\infty)$, there must exist some $\tilde{h}\in(\Delta,\infty)$ with $\varrho(\tilde{h})=\nu_2^2$. Clearly this $\tilde{h}$ satisfies equation (44) and $\tilde{h}>\Delta$. □

## Appendix E: Proof of Proposition 2

Note that for the triangular kernel, $\lambda_0=(2/3)\kappa_0$ and $\lambda_2=(2/5)\kappa_2$. By substituting these expressions into formula (30) for the efficiency ratio and cancelling out a common factor of $2\kappa_2/15$ from the numerator and denominator, it follows that

$$\text{Eff}_{\text{TS}}=\frac{5\kappa_0\kappa_2-15\phi(0)\psi(0)+3\phi^2(0)}{5\kappa_0\kappa_2-15\phi(\Delta)\psi(\Delta)+3\phi^2(\Delta)}\times\frac{(\kappa_0\kappa_2-\phi^2(\Delta))^2}{(\kappa_0\kappa_2-\phi^2(0))^2}. \tag{46}$$

Next $5\kappa_0\kappa_2-15\phi(\Delta)\psi(\Delta)+3\phi^2(\Delta)$ equals

$$\frac{5h^4}{24}-15\frac{h^4}{72}(1-3\delta^2+2\delta^3)(1-6\delta^2+8\delta^3-3\delta^4)+\frac{h^4}{12}(1-3\delta^2+2\delta^3)^2,$$

TABLE D.1
*Table of bandwidths that remove the leading order bias term, rounded to 4 decimal points in units of $\Delta$, for various kernel choices.*

| Kernel | Function $K(u)$ | Support | Bandwidth solving (44) |
|---|---|---|---|
| Boxcar | $1/2$ | $[-1,1]$ | $3.1342\Delta$ |
| Triangular | $(1-\lvert u\rvert)_+$ | $[-1,1]$ | $4.0343\Delta$ |
| Epanechnikov | $\frac{3}{4}(1-u^2)_+$ | $[-1,1]$ | $3.7866\Delta$ |
| Quartic | $\frac{15}{16}(1-u^2)_+^2$ | $[-1,1]$ | $4.3432\Delta$ |
| Triweight | $\frac{35}{32}(1-u^2)_+^3$ | $[-1,1]$ | $4.8367\Delta$ |
| Tricube | $\frac{70}{81}(1-\lvert u\rvert^3)_+^3$ | $[-1,1]$ | $4.3805\Delta$ |
| Cosine | $\frac{\pi}{4}\cos\left(\frac{\pi}{2}u\right)$ | $[-1,1]$ | $3.8613\Delta$ |

and so the first factor in (46) is

$$\frac{2}{5-5(1-3\delta^2+2\delta^3)(1-6\delta^2+8\delta^3-3\delta^4)+2(1-3\delta^2+2\delta^3)^2}.$$

Turning to the second factor

$$\kappa_0\kappa_2-\phi^2(\Delta)=\frac{h^4}{24}-\frac{h^4}{36}(1-3\delta^2+2\delta^3)^2=\frac{h^4}{72}\big(3-2(1-3\delta^2+2\delta^3)^2\big),$$

and so the second factor equals $(3-2(1-3\delta^2+2\delta^3)^2)^2$, establishing (33).

## Appendix F: Proof of Proposition 3

We want to show that this function

$$\mathrm{Eff}_{\mathrm{TS}}(\delta)=\frac{2\big(3-2(1-3\delta^2+2\delta^3)^2\big)^2}{5-5(1-3\delta^2+2\delta^3)(1-6\delta^2+8\delta^3-3\delta^4)+2(1-3\delta^2+2\delta^3)^2}$$

has a positive derivative for $0<\delta<1$. The numerator has degree 12 and the denominator has degree 7. The customary formula for the derivative of a rational function produces a rational function with a non-negative denominator and a numerator of degree 18. We will work through a sequence of steps reducing the degree of this polynomial to show that the numerator must be positive on $(0,1)$. That then rigorously establishes the monotonicity of $\mathrm{Eff}_{\mathrm{TS}}(\delta)$ which is visually apparent.

It is convenient to work instead with $x=1-\delta$. Then $1-3\delta^2+2\delta^3=3x^2-2x^3$ and $1-6\delta^2+8\delta^3-3\delta^4=4x^3-3x^4$. Therefore $\mathrm{Eff}_{\mathrm{TS}}(\delta)=f(1-\delta)$ where $f$ is a function given by

$$f(x)=\frac{2(3-2(3x^2-2x^3)^2)^2}{5-5(3x^2-2x^3)(4x^3-3x^4)+2(3x^2-2x^3)^2}$$

$$
\begin{aligned}
&= \frac{2(3-2x^4(3-2x)^2)^2}{5-5x^5(3-2x)(4-3x)+2x^4(3-2x)^2} \\
&= \frac{2(3-2x^4(3-2x)^2)^2}{5+[x^4(3-2x)][6-24x+15x^2]} \\
&= \frac{2(3-2g(x)(3-2x))^2}{5+g(x)(6-24x+15x^2)}
\end{aligned}
$$

for $g(x) = x^4(3-2x)$ and having replaced $\delta$ by $x = 1-\delta$ we will show that $f'(x) < 0$.

In the usual formula for the derivative of a ratio, $f'(x)$ has this numerator

$$
\begin{aligned}
n_1(x) = {} & 4(3-2g(x)(3-2x))(4g(x)-2g'(x)(3-2x))[5+g(x)(6-24x+15x^2)] \\
& -2(3-2g(x)(3-2x))^2[g'(x)(6-24x+15x^2)+g(x)(-24+30x)].
\end{aligned}
$$

Notice that $0 \leqslant g(x)(3-2x) \leqslant 1$ for $x \in [0,1]$ and so $3-2g(x)(3-2x) > 0$. As a result, the sign of $n_1(x)$ is preserved by dividing it by $2(3-2g(x)(3-2x))$, yielding

$$
\begin{aligned}
n_2(x) = {} & 2(4g(x)-2g'(x)(3-2x))[5+g(x)(6-24x+15x^2)] \\
& -(3-2g(x)(3-2x))[g'(x)(6-24x+15x^2)+g(x)(-24+30x)].
\end{aligned}
$$

Now since $g(x) = x^3[x(3-2x)]$ and $g'(x) = x^3[12-10x]$ and $x \in (0,1)$ we can divide $n_2(x)$ by $x^3/6$ finding that it has the same sign as

$$
\begin{aligned}
n_3(x) = {} & \frac{1}{3}\big(4x-2(12-10x)\big)(3-2x)[5+g(x)(6-24x+15x^2)] \\
& -\frac{1}{6}\big(3-2g(x)(3-2x)\big)\big[(12-10x)(6-24x+15x^2) \\
& \qquad +x(3-2x)(-24+30x)\big] \\
= {} & -8(1-x)(3-2x)[5+g(x)(6-24x+15x^2)] \\
& -(3-2g(x)(3-2x))[-35x^3+93x^2-70x+12] \\
= {} & -8(1-x)(3-2x)[5+g(x)(6-24x+15x^2)] \\
& -(3-2g(x)(3-2x))(1-x)(35x^2-58x+12).
\end{aligned}
$$

We can divide $n_3(x)$ by $-(1-x)$ getting a polynomial $n_4(x)$ with the opposite sign from $n_3$. This yields

$$
\begin{aligned}
& n_4(x) \\
& \quad = 8(3-2x)[5+g(x)(6-24x+15x^2)]+(3-2g(x)(3-2x))(35x^2-58x+12) \\
& \quad = 8(3-2x)[-30x^7+93x^6-84x^5+18x^4+5] \\
& \qquad +(-8x^6+24x^5-18x^4+3)(35x^2-58x+12) \\
& \quad = 8[60x^8-276x^7+447x^6-288x^5+54x^4-10x+15] \\
& \qquad +(-280x^8+1304x^7-2118x^6+1332x^5-216x^4+105x^2-174x+36)
\end{aligned}
$$

$$= 200x^8 - 904x^7 + 1458x^6 - 972x^5 + 216x^4 + 105x^2 - 254x + 156.$$

Note that the coefficient of $x^3$ in $n_4$ is a zero that must not be left out when entering the coefficients into symbolic differentiation codes.

We want to show that $n_4(x) > 0$ for $x \in (0,1)$ which then makes $n_3(x) < 0$ and ultimately $f'(x) < 0$ so that $\mathrm{Eff}'_{\mathrm{TS}}(\delta) > 0$ for $\delta = 1 - x \in (0,1)$.

Our next step is to show that $n_4''(x) > 0$ for all $x \in [0,1]$. Note that

$$n_4''(x) = 11200x^6 - 37968x^5 + 43740x^4 - 19440x^3 + 2592x^2 + 210.$$

Graphing $n_4''(x)$ versus $x$ and evaluating it numerically, makes it is clear that $148 < n_4''(x) < 369$. Our next few steps are just to eliminate even an unreasonable doubt about the sign of $n_4''$. Some readers might prefer to skip that and go instead to the subsection marked "Conclusion of the proof".

### *Positivity of $n_4''$*

We begin by writing

$$n_4'''(x) = 48x(1400x^4 - 3955x^3 + 3645x^2 - 1215x + 108).$$

Then by the triangle inequality,

$$|n_4'''(x)| \leqslant 48(1400 + 3955 + 3645 + 1215 + 108) = 495504 \leqslant 2^{19} \tag{47}$$

for all $x \in [0,1]$.

Now for $k \in \{0, 1, \ldots, 2^{12}\}$ define $x_k = k2^{-12}$. For each $k$ we will let $\widehat{n_4''(x_k)}$ be the numerical evaluation for the polynomial $n_4''(x_k)$ computed using Horner's method with double-precision in R, implemented with the 'horner' function in the **pracma** R package of Borchers (2019).

By formula (5.3) in Higham (2002), the absolute error in Horner's method for the polynomial $\sum_{r=0}^n a_r x^r$ is at most $\gamma_{2n}\tilde{p}(|x|)$ where $\gamma_{2n} \equiv 2nu/(1-2nu)$, $u$ is the unit roundoff, and $\tilde{p}(|x|) = \sum_{r=0}^n |a_r||x|^r$.

Applying that bound to our 6th degree polynomial $n_4''(x)$, and noting that for each $x \in [0,1]$, $\tilde{p}(|x|)$ will be at most the sum of the absolute value of the coefficients we find that

$$\max_{0 \leqslant k \leqslant 2^{12}} \left| n_4''(x_k) - \widehat{n_4''(x_k)} \right| \leqslant \gamma_{2\times 6}\tilde{p}(|x_k|) \leqslant \frac{12u}{1-12u} \times 115{,}150. \tag{48}$$

We need not worry about floating point error induced by evaluating $x_k$, because each $x_k$ is a floating point number. For double precision in R, the unit roundoff is $u = 2^{-53} \leqslant 2\times 10^{-15}$ from which $12u/(1-12u) \leqslant 10^{-13}$, and so the maximum error in (48) is at most $10^{-7}$. The smallest value of $\widehat{n_4''(x_k)}$ among all $2^{12}+1$ evaluation points $x_k$ was 148.5743 and so $\min_{0\leqslant k\leqslant 2^{12}} n_4''(x_k) \geqslant 148$.

Now we are ready to prove that $n_4''(x) > 0$ for all $x \in [0,1]$. For any $x \in [0,1]$ there exists $k_* \in \{0, 1, \ldots, 2^{12}\}$ with $|x - x_{k_*}| \leqslant 2^{-13}$. By (47) we know that $n_4''$ is Lipschitz continuous on $[0,1]$ with Lipschitz constant $2^{19}$. Therefore

$$n_4''(x) \geqslant n_4''(x_{k_*}) - |n_4''(x_{k_*}) - n_4''(x)| \geqslant 148 - 2^{19} \times 2^{-13} > 0,$$

holds for all $x \in [0,1]$.

***Conclusion of the proof***

We have shown above that $n_4''(x) > 0$ for any $x \in [0,1]$. Now since $n_4'(1) = -20$, and $n_4''(x) > 0$ for all $x \in [0,1]$, it follows that $n_4'(x) < 0$ for all $x \in [0,1]$. Finally since $n_4(1) = 5$ and $n_4'(x) < 0$ for all $x \in [0,1]$, it follows that $n_4(x) > 0$ for all $x \in [0,1]$. This completes the proof. □

## Acknowledgments

We thank Hal Varian and Harrison Li for commenting on the paper as well as Steve Marron and Wolfgang Härdle for some discussions about nonparametric regression. We also thank anonymous reviewers for comments that led us to improve the paper.

## References

[1] Abdulkadiroğlu, A., Angrist, J. D., Narita, Y. and Pathak, P. (2022). Breaking ties: Regression discontinuity design meets market design. *Econometrica* **90** 117–151.

[2] Angrist, J., Autor, D. and Pallais, A. (2020). MARGINAL EFFECTS OF MERIT AID FOR LOW-INCOME STUDENTS Technical Report, National Bureau of Economic Research.

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

[4] Angrist, J. D. and Lavy, V. (2009). Replication data for: Using Maimonides' Rule to Estimate the Effect of Class Size on Student Achievement.

[5] Angrist, J. D., Lavy, V., Leder-Luis, J. and Shany, A. (2019). Maimonides' Rule Redux. *American Economic Review: Insights* **1** 309–24.

[6] Borchers, H. W. (2019). pracma: Practical Numerical Math Functions R package version 2.2.9.

[7] Boruch, R. F. (1975). Coupling randomized experiments and approximations to experiments in social program evaluation. *Sociological Methods & Research* **4** 31–53.

[8] Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica* **82** 2295–2326.

[9] Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2018). On the Effect of Bias Estimation on Coverage Accuracy in Nonparametric Inference. *Journal of the American Statistical Association* **113** 767-779.

[10] Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2019). Optimal bandwidth choice for robust bias-corrected inference in regression discontinuity designs. *The Econometrics Journal* **23** 192-210.

[11] Campbell, D. T. (1969). Reforms as experiments. *American psychologist* **24** 409.

[12] Cappelleri, J. C., Darlington, R. B. and Trochim, W. M. K. (1994). Power Analysis of Cutoff-Based Randomized Clinical Trials. *Evaluation Review* **18** 141-152.
[13] Cappelleri, J. C. and Trochim, W. M. K. (2003). Cutoff Designs. In *Encyclopedia of Biopharmaceutical Statistics* (M. Dekker, ed.) CRC Press.
[14] Cattaneo, M. D., Frandsen, B. R. and Titiunik, R. (2015). Randomization inference in the regression discontinuity design: An application to party advantages in the US Senate. *Journal of Causal Inference* **3** 1–24.
[15] Cattaneo, M. D. and Titiunik, R. (2022). Regression discontinuity designs. *Annual Review of Economics* **14** 821–851.
[16] Cheng, M.-Y., Fan, J. and Marron, J. S. (1997). On automatic boundary corrections. *The Annals of Statistics* **25** 1691–1708.
[17] Crump, R. K., Hotz, V. J., Imbens, G. W. and Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96** 187–199.
[18] Eckles, D., Ignatiadis, N., Wager, S. and Wu, H. (2020). Noise-induced randomization in regression discontinuity designs Technical Report, arXiv:2004.09458.
[19] Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications: monographs on statistics and applied probability 66* **66**. CRC Press, Boca Raton, FL. MR1383587
[20] Gelman, A. and Imbens, G. (2019). Why High-Order Polynomials Should Not Be Used in Regression Discontinuity Designs. *Journal of Business & Economic Statistics* **37** 447–456.
[21] Goldberger, A. S. (1972). Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations Technical Report No. Discussion paper 128–72, Institute for Research on Poverty, University of Wisconsin–Madison.
[22] Hahn, J., Todd, P. and der Klaauw, W. V. (2001). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica* **69** 201–209.
[23] Higham, N. J. (2002). *Accuracy and Stability of Numerical Algorithms*, Second ed. Society for Industrial and Applied Mathematics, Philadelphia.
[24] Imbens, G. and Kalyanaraman, K. (2012). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies* **79** 933–959.
[25] Jacob, R., Zhu, P., Somers, M.-A. and Bloom, H. (2012). A Practical Guide to Regression Discontinuity. *MDRC*.
[26] Krantz, C. (2022). Modeling the outcomes of a longitudinal tie-breaker regression discontinuity design to assess an in-home training program for families at risk of child abuse and neglect, PhD thesis, University of Oklahoma.
[27] Li, H. H. and Owen, A. B. (2022). A general characterization of optimal tie-breaker designs Technical Report No. arXiv:2202.12511, Stanford University.
[28] Metelkina, A. and Pronzato, L. (2017). Information-regret compro-

mise in covariate-adaptive treatment allocation. *The Annals of Statistics* **45** 2046–2073.

[29] Morrison, T. P. and Owen, A. B. (2022). Optimality in Multivariate Tie-breaker Designs Technical Report No. arXiv:2202.10030, Stanford University.

[30] Owen, A. B. and Varian, H. (2020). Optimizing the tie-breaker regression discontinuity design. *Electronic Journal of Statistics* **14** 4004–4027.

[31] Rosenman, E. and Rajkumar, K. (2019). Optimized Partial Identification Bounds for Regression Discontinuity Designs with Manipulation Technical Report No. arXiv:1910.02170, Stanford University.

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

[33] Wand, M. P. and Jones, M. C. (1994). *Kernel smoothing*. Chapman Hall/CRC, Boca Raton, FL.