NORTHWESTERN UNIVERSITY

Essays in Microeconometrics

A DISSERTATION

SUBMITTED TO THE GRADUATE SCHOOL

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

for the degree

DOCTOR OF PHILOSOPHY

Field of Economics

By

Vishal Kamat

EVANSTON, ILLINOIS

June 2018

Abstract

Essays in Microeconometrics

Vishal Kamat

This dissertation consists of three chapters in microeconometrics. Each of the chapters corresponds to a paper that studies topics related to program and treatment effects, which has been the subject of research in econometric theory and empirical applied microeconomics.

The first chapter is a paper that studies identification of program effects in settings with latent choice sets. Here, by latent choice sets, I mean the unobserved heterogeneity that arises when the choice set from which the agent selects treatment is heterogeneous and unobserved by the researcher. The analysis is developed in the context of the Head Start Impact Study, a social experiment designed to evaluate preschools as part of Head Start, the largest early childhood education program in the United States. In this setting, resource constraints limit preschool slots to only a few eligible children through an assignment mechanism that is not observed in the data, which in turn introduces unobserved heterogeneity in the child's choice set of care settings. I propose a nonparametric model that explicitly accounts for latent choice sets in the care setting enrollment decision. In

this model, I study various parameters that evaluate Head Start in terms of policies that mandate enrollment and also those that allow voluntary enrollment into Head Start. I show that the identified set for these parameters given the information provided by the study and by various institutional details of the setting can be constructed using a linear programming method. Applying the developed analysis, I find that a significant proportion of parents voluntarily enroll their children into Head Start if provided access and that Head Start is effective in terms of improving short-term test scores across multiple policy dimensions.

The second chapter is a paper that is joint work with Ivan A. Canay and is forthcoming in the *The Review of Economic Studies* - see Canay and Kamat (2017). Here we study a question of statistical testing in the regression discontinuity design. In the regression discontinuity design, it is common practice to assess the credibility of the design by testing whether the means of baseline covariates do not change at the cutoff (or threshold) of the running variable. This practice is partly motivated by the *stronger* implication derived by Lee (2008), who showed that under certain conditions the distribution of baseline covariates in the RDD must be continuous at the cutoff. We propose a permutation test based on the so-called induced ordered statistics for the null hypothesis of continuity of the distribution of baseline covariates at the cutoff; and introduce a novel asymptotic framework to analyze its properties. The asymptotic framework is intended to approximate a *small* sample phenomenon: even though the total number n of observations may be large, the number of *effective* observations local to the cutoff is often small. Thus, while traditional asymptotics in RDD require a growing number of observations local to the cutoff as $n \to \infty$, our framework keeps the number q of observations local to the cutoff

fixed as $n \to \infty$. The new test is easy to implement, asymptotically valid under weak conditions, exhibits finite sample validity under stronger conditions than those needed for its asymptotic validity, and has favorable power properties relative to tests based on means. In a simulation study, we find that the new test controls size remarkably well across designs. We then use our test to evaluate the plausibility of the design in Lee (2008), a well-known application of the RDD to study incumbency advantage.

The third chapter is a paper that is forthcoming in *Econometric Theory* - see Kamat (2017). Here I study the validity of nonparametric tests used in the regression discontinuity design. The null hypothesis of interest is that the average treatment effect at the threshold in the so-called sharp design equals a pre-specified value. I first show that, under assumptions used in the majority of the literature, for *any* test the power against any alternative is bounded above by its size. This result implies that, under these assumptions, any test with nontrivial power will exhibit size distortions. I next provide a sufficient strengthening of the standard assumptions under which I show that a version of a test suggested in Calonico et al. (2014a) can control limiting size.

Acknowledgements

I am greatly indebted to my main advisor Ivan Canay, whose generous guidance and support beginning from the summer of my first year has crucially benefited me during my time at Northwestern. I am also extremely grateful to my other advisors Chuck Manski and Alex Torgovitsky who along with Ivan have been a continuous source of inspiration and learning. Their ideas and approach to econometrics have greatly influenced mine and especially those that are presented in this dissertation. Many other professors and friends in the supportive environment of the Economics department at Northwestern have also played an important role in the development of the ideas presented in this dissertation.

Finally, I would like to give special thanks to some for their encouragement, love and tremendous emotional support they have provided me during the last five years. My parents Medha and Anand for believing and supporting me in my education, my brother Viraj for inspiring me to start a PhD, and Lola without whom I would not have been able to successfully finish it.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Identification with Latent Choice Sets: The Case of the Head Start Impact Study

1.1. Introduction

This paper studies identification of program effects in settings with latent choice sets. Here, by latent choice sets, I mean the unobserved heterogeneity that arises when the choice set from which the agent selects treatment is heterogeneous and unobserved by the researcher. Such unobserved heterogeneity in choice sets is common to data on public programs, where resource constraints limit program access to only a few eligible agents through an assignment mechanism that is not observed by the researcher. In these settings, the agent's decision to select into treatments is based not only on the preferences over all the treatments but also on the choice set of available treatment options. In this paper, I propose a nonparametric model that explicitly accounts for latent choice sets in the treatment selection decision and study the identification of various program effect parameters in the context of this model.

The analysis in this paper is developed in the context of the Head Start Impact Study (HSIS), which was a social experiment designed to evaluate Head Start. Head Start is the largest early childhood education program in the United States, which provides free preschool education to three- and four-year-old children from low income eligible households. As noted previously in Kline and Walters (2016), preschools are resource

constrained to the extent that available slots are limited to only a few eligible children. In this setting, the choice set of care settings that the parents face for their child is determined by their applications to various preschools and by how these preschools allocate slots. However, the HSIS does not provide data on these decisions leading to the parents' possibly constrained choice set of care settings. Instead, it only provides data on the resulting care setting selected from the given choice set. Despite this, as further discussed in Section 1.3.1, previous studies on the HSIS employ choice models that do not account for the unobserved heterogeneity that may be present in the choice sets, and implicitly assume all preschools are available to every child.

This paper develops a nonparametric framework that aims to explicitly account for the latent choice set of care settings in the care setting selection decision. The proposed model treats the observed selected care setting as simply the product of the parents' utility maximization decision given the parents' choice set of care settings. In particular, the model is entirely nonparametric and only assumes that parents behave rationally and have a strict preference relationship over all care settings. Importantly, the parents' obtained choice set is permitted to include only a subset of preschools and to be correlated in an unrestricted manner with their preferences and their child's potential outcomes under the various care settings. Moreover, by treating this choice set as a latent variable, the model assumes that the researcher does not observe the parents' obtained choice set. As further discussed in Remark 1.3.1, the proposed model builds on the discrete choice framework of Manski (2007) by allowing for latent choice sets of care settings and potential outcomes associated with each care setting.

In the proposed model, I study a range of parameters that evaluates Head Start in comparison to alternate care settings across multiple policy dimensions. The first class of parameters evaluates the effect on a child's short-term test score outcome of Head Start in comparison to home care, i.e. no preschool. This class of parameters corresponds to evaluating policies that mandate Head Start enrollment versus no preschool enrollment. As noted in early work by Heckman et al. (1997) and Manski (1996, 1997a), many policies however do not mandate attendance but rather allow agents to voluntarily select into their preferred treatment option. In turn, I also study a class of parameters that evaluates the effect of providing Head Start access to parents. That is a Head Start option from which the parents can voluntarily choose whether to enroll their child.

The identification analysis begins by studying what we can learn about the parameters of interest using only the restrictions imposed on the model by the HSIS experiment. These restrictions correspond to those that the observed data and the experimental design of the HSIS place on the underlying model distribution. The parameters of interest are in general not point identified but rather partially identified. Deriving the identified set analytically for these parameters is difficult due to the complicated structure of the model and the imposed restrictions. Instead, I propose a general computational method to obtain the identified set for the various parameters given restrictions imposed on the model. For computational tractability, the proposed method requires that the underlying model variables are discrete, or transformed into discrete variables, and that the imposed restrictions are linear in the model distribution. Under these requirements, I show that the identified set can be computed by solving linear programming problems. I illustrate that the restrictions imposed by the HSIS experiment satisfy these requirements. An

important benefit of the computational method is the flexibility by which restrictions imposed by additional assumptions on the model can also be studied. With the aim of obtaining stronger conclusions, I discuss several additional nonparametric assumptions motivated by unique institutional details of the HSIS setting.

The developed identification analysis is applied using the empirical distribution of the HSIS sample data to estimate the identified sets for the parameters of interest. In order to construct confidence intervals for these parameters, I discuss how the profiled subsampling method proposed by Romano and Shaikh (2008a) can be applied in a computationally tractable manner. The estimated identified sets are informative. For example, using only the information provided by the experimental design of the HSIS, I find that between 79.9% to 91.7% of parents prefer to enroll their child into Head Start when they do not have any alternate preschool option. Amongst this subgroup of parents who prefer to enroll their child, I find under an additional nonparametric monotonicity assumption and a nonparametric assumption on how the experiment affects the choice set of care settings that between 4.9% to 42.7% of their children strictly benefit in terms of improving their short-term test scores. Furthermore, these estimated identified sets are tighter for some specific subgroups and under additional nonparametric assumptions. In summary, the findings suggest towards the benefits of Head Start and qualitatively corroborate those of previous studies on the HSIS under weak nonparametric assumptions.

The remainder of the paper is organized as follows. Section 1.2 describes the experimental design of the HSIS, specifically focusing on how the parents may face heterogeneous choice sets of care settings. Based on this description, Section 1.3 introduces the formal model used to analyze the HSIS and also provides a comparison to previously used models

in this setting. Section 1.4 illustrates various parameters of interest and the formal identification analysis. Section 1.5 discusses performing statistical inference on the parameter of interest using the HSIS sample data. Section 1.6 presents the empirical results. Here I first present results using only the restrictions imposed by the HSIS experiment and then under restrictions imposed by additional nonparametric assumptions motivated by the details of the HSIS setting. Section 1.7 concludes.

1.2. Experimental Design of the HSIS

Head Start is the largest early childhood education program in the United States. The program provides free preschool education to three- and four-year-old children from disadvantaged households. Program eligibility for households is primarily determined by the federal poverty line, yet certain exceptions qualify additional low income households. As part of a congressional mandate, the Head Start Impact Study (HSIS) was a social experiment implemented in the beginning of Fall 2002 with the aim of evaluating the impact of the program.

The experiment acquired a sample of Head Start preschool centers and participating children using a multistage stratified sampling scheme, an extensive description of which is provided in Puma et al. (2010). The centers and the children were sampled not from the entire population, but rather from specific sub-populations. More specifically, centers were first sampled from a so-called saturated population of Head Start centers, which referred to centers where the number of available slots was strictly smaller than the number of applicants. From each sampled center, children were then sampled separately from the sub-population of newly entering three- and four-year-old eligible applicants in Fall 2002,

where newly entering referred to children who were not previously enrolled in any Head Start services. As noted in Puma et al. (2010), the sample of three-year-old children differed considerably from that of four-year-old children in terms of observed covariates. Following previous studies, the analysis in this paper is hence separately performed for the two age samples. Note that the sub-population restrictions introduced on the selected sample may raise possible concerns on the external validity of the empirical results using this sample.

Next, I provide a brief description of the HSIS experimental design. In order to clearly emphasize the aspects of this design that this paper focuses on, I organize this description into the following stages:

> **Stage 1**: At each sampled center, the experiment randomly offered center access to the children sampled from that center. To be precise, let Z be an indicator denoting the randomized center offer to a given child. If $Z = 1$, the child was granted access to that center for two consecutive years if the child was a three-year-old and for a single year if the child was a four-year-old. Whereas, if $Z = 0$, the child's parents were told access to that center would not be granted for that year, but if the child was a three-year-old the parents could choose to reapply for access in the following year, i.e. Fall 2003. This asymmetry in the experimental offer for the two age groups naturally arose as Head Start services are unavailable to five year old children. However, the experiment did not control the child's chances of obtaining an offer from other preschools, both Head Start centers from the one the child was sampled from and alternate non Head Start preschools. In particular, the child's parents could choose to apply to these other preschools

and in turn possibly obtain access, which primarily depended on whether the corresponding applied preschools too were saturated or not. As a fallback, the parents also always had the option to care for their child at home by themselves, a relative or some known individual.

Stage 2a: After obtaining their respective choice set of care settings based on the experimental offer and their application decisions to other preschools, the parents then decided in which care setting to enroll their child for in Fall 2002.

Stage 2b: The experiment further allowed the parents' enrollment decision to vary over the school year across different care settings, i.e. the child could attend multiple care settings for varying duration between Fall 2002 and Spring 2003.

Stage 3: After the child completed an entire school year enrolled in various selected care settings, the experiment collected data on a number of outcomes in Spring 2003. The experiment further continued to collect outcomes for upto four additional years in the child's corresponding care setting for those years.

For the purposes of this paper, it is most important to note that the experiment did not provide any information on the parents' application decisions to other preschools in Stage 1, which implies that we do not observe the parents' choice set. To the extent that the parents' preferred care setting may not have been available in their choice set, it is important to account for this fact when studying the parents' enrollment decisions in Stage 2a. The aim of this paper is to develop a framework that accounts for this feature of the experiment when analyzing the experimental data.

Note that certain aspects of the experiment involved important dynamic features. In particular, Stage 2b introduced a dynamic decision process and a duration treatment, and

Table 1.1. Percentage in each care setting by offer group and by each age group.

	$Z=1$				$Z=0$			
	n_1	Head Start	Alternate	Home care	n_0	Head Start	Alternate	Home care
Age three	1188	86.62	5.13	8.25	678	14.16	25.96	59.88
Age four	964	79.98	11.20	8.82	557	11.13	38.78	50.09

Notes: n_1 denotes the number of children with randomized offer, and n_0 those without randomized offer. Appendix A.0.3 provides details on how the data used for these statistics was constructed.

Stage 3 introduced a panel of outcomes under differing care settings every year. However, the formal model proposed in the following section is static in the sense that it does not account for either of these dynamic features. Instead, following previous studies on the HSIS, the analysis in this paper uses an administratively coded focal care setting as the unique enrollment decision for the entire year, i.e. it essentially combines Stage 2a and Stage 2b - see Table 1.1 for some descriptive statistics on this enrollment decision. Moreover, following Kline and Walters (2016), the analysis in this paper considers a single outcome of interest taken to be the average of the Woodcock Johnson III (WJIII) test score and the Peabody Picture and Vocabulary Test (PPVT) test score in Spring 2003, which is then standardized with respect to the corresponding baseline test score average in Fall 2002.

Furthermore, note that the above description had no mention of observed covariates with respect to sampled Head Start centers and sampled children. Bloom and Weiland (2015) and Walters (2015) have previously noted that Head Start center covariates are related to the variation in child outcomes across centers. In order to account for this in the empirical analysis, I focus on two covariates that have been regarded as important measures of preschool quality: (i) HC denotes an indicator for whether the Head Start center had the HighScope curriculum, which was part of the influential Perry Preschool

program; and (ii) CS denotes an indicator for whether the Head Start center had a low class size ratio.

1.3. Model Framework

In this section, I propose a nonparametric model that generates the parents' observed selected care setting for their child and the child's observed test score outcome under the selected care setting. The proposed model is tightly connected to the various stages of the experimental design presented in the previous section. Before proceeding to these stages, following the analysis in Kline and Walters (2016), I begin by assuming that the set of all possible care settings, i.e. the set of treatments, is categorized into the following

$$\mathcal{D} = \{0, 1, 2\} \ ,$$

where 0 denotes home care or no preschool, 1 denotes an alternate or non Head Start preschool and 2 denotes a Head Start preschool.

In Stage 1, parents apply to both Head Start and alternative preschools and as a product obtain their choice set of care settings. This choice set implicitly encompasses the costs that parents face to acquire a preschool in their choice set. Let $C(1)$ denote the parents' (potential) choice set had access been granted to the Head Start center as part of the experiment, and let $C(0)$ denote the parents' (potential) choice set had access not been granted. As parents always had the fallback option of home care, these choice sets are assumed to possibly take values in the following set

$$\mathcal{C} = \{\{0\}, \{0, 1\}, \{0, 2\}, \{0, 1, 2\}\} \ ,$$

i.e. the set of all possible choice sets that contain home care. Let C denote the parents' obtained choice set, which is given by the following relationship

$$C = C(1)Z + C(0)(1 - Z) \ .$$

In Stage 2, parents decide in which care setting to enroll their child given their choice set obtained from Stage 1. Let $U(d)$ denote the indirect utility that the parents would obtain had their child been enrolled in care setting $d \in \mathcal{D}$. In particular, this utility corresponds to the parents' benefits and costs that they face if their child was enrolled in a given care setting. Let D denote the observed enrollment decision, which is assumed to be given by the following utility maximization relationship

$$D = \arg\max_{d \in C} U(d) \ . \tag{1.1}$$

Note that since the utility under each care setting does not possess any cardinal value, different monotonic transformations of these utilities will generate observationally equivalent choices. For the purposes of the analysis, it is hence more useful to directly refer to the parents' underlying preference type that these utilities represent. To this end, assume that each parent has a strict preference relation over the set of care settings, which then implies that the utilities can be strictly ordered, i.e.

$$U_{(0)} > U_{(1)} > U_{(2)}$$

where $U_{(k)}$ denotes the $(k+1)$th largest value of the utilities $\{U(d) : d \in \mathcal{D}\}$. Denoting by d_k the care setting in $\mathcal{D}$ with the $(k+1)$th largest utility for the parents, i.e. $U_{(k)} = U(d_k)$,

the parents' underlying preference type is then simply given by the following label

$$U = (d_0 \succ d_1 \succ d_2) \ ,$$

which corresponds to the preference ordering on $\mathcal{D}$ that the utilities $\{U(d) : d \in \mathcal{D}\}$ represent. Note that since there are three possible care settings, this implies that there are six possible preference orderings over them. To be concrete, this set can explicitly be stated as follows

$$\mathcal{U} = \{(2 \succ 1 \succ 0),\ (2 \succ 0 \succ 1),\ (1 \succ 2 \succ 0),\ (1 \succ 0 \succ 2),\ (0 \succ 2 \succ 1),\ (0 \succ 1 \succ 2)\} \ .$$

Using the above notation, the utility maximization relationship in (1.1) can be re-written in terms of the preference types and the obtained choice set through the following relationship

$$D = \sum_{u \in \mathcal{U}, c \in \mathcal{C}} d_{u,c} I\{U = u, C = c\} \equiv d_{U,C} \ , \tag{1.2}$$

where

$$d_{U,c} = \arg\max_{d \in c} U(d)$$

denotes the preferred care setting for the preference type U under a non-empty subset $c \subseteq \mathcal{D}$, i.e. under a given possible choice set.

In Stage 3, the child's observed test score outcome is realized under the enrolled care setting from Stage 2. Let $Y(d)$ denote the (potential) test score outcome had the child been enrolled in care setting $d \in \mathcal{D}$. Let Y denote the observed test score outcome under

the enrolled care setting. The observed test score is related to the potential test scores and the enrolled care setting through the following relationship

$$Y = \sum_{d \in \mathcal{D}} Y(d) I\{D = d\} \equiv Y(D) \ . \tag{1.3}$$

Next, I emphasize two important aspects of the above described model. First, the model placed no restrictions on the dependence between the underlying variables in the three model stages, which implies that the potential outcomes, the preference type and the choice sets are allowed to be statistically dependent. As a result, this allows the obtained choice sets to be dependent on the preference type and the potential test scores, and further the selected care setting to be dependent on the potential outcomes.

Second, as the experiment did not provide data on the parents' application decisions, the value of the obtained choice set in Stage 1 of the model is not observed. Recall however that the experimental design randomly offered Head Start access to children, which implies that Head Start would be present in the obtained choice set had an offer been made. This partial information formally imposes some restrictions on the model, which are stated in the form of an assumption below.

Assumption HSIS. *Let the underlying model variables be such that*

(i) $2 \in C(1)$.

(ii) $(Y(0), Y(1), Y(2), U, C(0), C(1)) \perp Z \mid (HC, CS)$.

Assumptions HSIS(i) states that being assigned an offer guaranteed that a Head Start preschool was in the parents' choice set. To be concrete, it ensures that the choice set

$C(1)$ can possibly only take values in the following set

$$\mathcal{C}_1 = \{\{0,2\},\{0,1,2\}\} \ .$$

Assumption HSIS(ii) states that conditional on the Head Start center, access to that center was randomly assigned to the children.

Remark 1.3.1. In the context of observed choice data, using an agent's underlying preference type U was first proposed by Marschak et al. (1959) for the purposes of testing an agent's rationality. Manski (2007) extended this idea to propose a general nonparametric discrete choice framework by formally introducing the selection equation in (1.2). In the setting of observed and statistically independent choice sets, Manski (2007) then developed the analysis to predict choice probabilities under various counterfactual choice sets. See also Kline and Tartari (2016) and Manski (2014) for applications of this framework to study labor supply decisions. The model discussed in this section builds on that of Manski (2007) in two directions essential to the experimental design of the HSIS. First, the model allows for the obtained choice set in Stage 1 to be unobserved by the researcher and also to be correlated with the preferences. Second, the model introduces a Stage 3 with the outcome equation in (1.3), where the selection equation in (1.2) is an intermediate step. ■

1.3.1. Previous Models on the HSIS

Kline and Walters (2016) have previously proposed a parametric structural model to evaluate the HSIS. The most important distinction between their choice model and that proposed in the previous section is the conceptual differences in the description of the

parents' decision process. In particular, their choice model assumes that the utility from enrolling in the various care settings is given by

$$U(2) \equiv U(2, Z),\ U(1)\ ,\ \text{and}\ U(0)\ ,$$

i.e. the randomized offer affects the utility under Head Start, and that the selected care setting is then given by the following relationship

$$D = \arg\max_{d \in \mathcal{D}} U(d)\ ,$$

i.e. utility is maximized over all possible care settings. By imposing that parents maximize enrollment utility from the set of all care settings, this choice model implicitly assumes that parents face all care settings when making enrollment decisions. More specifically, it does not explicitly account for Stage 1 of the experimental setting, i.e. the fact that parents may not face all preschools in their choice set when making enrollment decisions. As a result, the underlying choice model may possibly mis-specify the choice set of obtained care settings - see Stopher (1980) and Williams and Ortúzar (1982) for early criticisms on mis-specifying choice sets in parametric discrete choice models.

In contrast, their choice model implicitly combines Stage 1 and Stage 2 into a single stage. To be specific, the choice model does not treat the utility as solely the parents' utility from enrolling their child in a given care setting in Stage 2, i.e. the parents' benefits and costs of enrolling their child in a given care setting. Instead, the utility implicitly also encompasses the costs from Stage 1 that the parents face to acquire a preschool in their choice set, as noted by how the experimental offer affects the utility of enrolling in a Head Start preschool. This treatment of the utility may however not capture the

structure present in the decision problem of the parents. Moreover, it does not permit the study of counterfactuals that relate to the parents' preferences over the care settings in Stage 2.

A further distinction between the model in Kline and Walters (2016) and that proposed in the previous section are the various additional assumptions imposed. In particular, their model imposes assumptions such as separability and parametric specifications on both the utilities and potential outcomes. Unlike, for example, Assumption HSIS, such assumptions are not motivated by nonparametric arguments from the empirical setting. Instead, these arguments are based on conditions required to point identify different parameters in the model. This is in contrast to the partial identification direction pursued in this paper, which is valid under weaker assumptions.

Remark 1.3.2. Walters (2015) also proposes a parametric structural model to evaluate the variation in parameters across different Head Start centers sampled as part of the HSIS. This model focuses on a setting with only two treatments consisting of a Head Start preschool and no Head Start preschool, i.e. alternative preschools and home care are grouped into a single treatment. The discussion in this section also conceptually applies to this model. ■

Remark 1.3.3. Previous work on discrete choice models has proposed parametric approaches that allow for latent choice sets such as, for example, Ben-Akiva and Boccara (1995), which also provides a discussion on the various approaches. These approaches in essence augment the standard parametric discrete choice models with a parametric

mixture model over the unobserved choice sets. The approach pursued in this paper is instead nonparametric and is conceptually very different. ■

1.4. Identification Analysis

In this section, I first describe various parameters of interest based on the model proposed in the previous section and then develop the subsequent identification analysis. In particular, as further discussed in Remark 1.4.3 below, I note that the developed identification analysis is based on a general finite dimensional computational approach, which specifically requires all the observed and underlying variables used in the analysis to be discrete in nature. For the model described for the HSIS in the previous section, this is the case except for the test score outcome of interest. Hence, for the purposes of the analysis, I take a simple binary transformation of the test score outcome that corresponds to an intuitive summary indicator to evaluate whether a given test score is viewed as high or low. To be formal, denote by

$$Y^{\dagger} \equiv 1\{Y > 0\} \tag{1.4}$$

a binary transformation of the observed test score outcome, and, similarly, denote by

$$Y^{\dagger}(d) \equiv 1\{Y(d) > 0\} \tag{1.5}$$

a binary transformation of the potential test score outcome under each care setting $d \in \mathcal{D}$. Recall that the test scores are standardized with respect to the corresponding baseline test scores in Fall 2002. In turn, under these binary transformations, a value of one signifies a "high" test score in the sense of being above the baseline mean test score and a value of

zero signifies a "low" test score in the sense of being below the baseline mean test score. However, note that in principle alternate discrete transformations of the outcomes are also permitted by the identification analysis.

In Section 1.4.1 below, I begin by describing parameters of interest that are formally defined as functions of the model distribution under the transformed binary outcomes. In order to be formal in the description of these parameter, I first introduce some additional notation here. To this end, denote by

$$W = (Y^{\dagger}(0), Y^{\dagger}(1), Y^{\dagger}(2), U, C(0), C(1), Z, HC, CS) \sim Q \tag{1.6}$$

the random variable that summarizes the underlying model variables under the transformed potential test scores for each child, where Q denotes the distribution of this random variable defined on a discrete sample space $\mathcal{W} = \{0,1\}^3 \times \mathcal{U} \times \mathcal{C} \times \mathcal{C} \times \{0,1\}^3$. Due to the discreteness, note that Q is a probability mass function with support contained in $\mathcal{W}$, i.e. $Q : \mathcal{W} \to [0,1]$ such that

$$\sum_{w \in \mathcal{W}} Q(w) = 1 \ .$$

Given a parameter of interest based on the unknown probability mass function Q, I formally state in Section 1.4.2 below the identification problem, i.e. what we can learn about the parameter given the known distribution of the observed data and the assumptions imposed on the model. Similar to the above notation, denote by

$$X = (Y^{\dagger}, D, Z, HC, CS) \sim P \tag{1.7}$$

the random variable that summarizes the observed random variables under the transformed observed outcome for each child, where P denotes the distribution of this random variable defined on a discrete sample space $\mathcal{X} = \{0,1\} \times \mathcal{D} \times \{0,1\}^3$. Similar to Q, note that P is a probability mass function with support contained in $\mathcal{X}$, i.e. $P : \mathcal{X} \to [0,1]$ such that

$$\sum_{x \in \mathcal{X}} P(x) = 1 \ .$$

Finally, I show in Section 1.4.3 below that what we can learn about the parameter of interest can be stated as solutions to optimization problems. Before proceeding, I note that the various conditions required to ensure that these optimization problems can be restated as tractable linear programs motivate in part the choice of parameters and assumptions studied in Section 1.4.1 and Section 1.4.2. For expositional reasons, in these sections below, I only make note of these choices when they are formally presented as assumptions and leave the discussion on their importance in terms of computation after I present the identification result in Proposition 1.4.1 in Section 1.4.3. In these sections below, I repeatedly use w to refer to

$$(y(0), y(1), y(2), u, c(0), c(1), z, hc, cs) \in \mathcal{W} \ ,$$

i.e. a generic value in the sample space of the underlying random variables.

1.4.1. Parameters of Interest

In the proposed model with the transformed outcomes, one could possibly consider a number of interesting parameters with the aim of evaluating Head Start. The developed

identification analysis allows for the flexibility to study a number of such parameters as long as they satisfy a specific condition. As discussed in Section 1.4.3, this condition is motivated in part by those required to ensure that the final optimization problem is a linear program. I formally state this condition in the following assumption.

Assumption 1.4.1. *The parameter of interest $\theta(Q)$ can be written as*

$$\theta(Q) = \frac{\sum_{w \in \mathcal{W}} a_{num}(w) \cdot Q(w)}{\sum_{w \in \mathcal{W}} a_{den}(w) \cdot Q(w)} \ , \tag{1.8}$$

where $a_{num}, a_{den} : \mathcal{W} \to \mathbf{R}$ are known functions.

Assumption 1.4.1 states that the parameters of interest are required to be fractions of linear functions of the probability mass function of the model random variables. Note that a special case of this class of functions are linear functions, where by construction the denominator takes a value of one. I next describe several interesting parameters that I study in the empirical results, which satisfy this assumption and also correspond to policies evaluating Head Start across multiple dimensions. These policies in particular correspond to those that mandate Head Start enrollment and also those that allow voluntary enrollment into Head Start - see Heckman et al. (1997) and Manski (1996, 1997a) for early studies, and also Heckman et al. (2006, 2008) and Heckman and Vytlacil (2007) for examples of more recent studies evaluating the latter policies.

Note however that Assumption 1.4.1 is flexible in the sense that it allows a researcher to also choose and study parameters that correspond to other policy dimensions. In order to emphasize this flexibility, I state below the class of parameters that I study in the form

of examples of Assumption 1.4.1. In these examples below, I denote by

$$D_c = \sum_{u \in \mathcal{U}} d_{u,c} I\{U = u\}$$

the care setting in which the parent would enroll their child had their choice set of care settings been exogenously pre-specified to a set $c \subseteq \mathcal{D}$.

Example 1.4.1. *(Head Start versus Home care)* This class of parameters aims to evaluate Head Start in comparison to home care, which corresponds to comparing policies mandating Head Start enrollment versus mandating home care. I begin by considering

$$\text{PB}_{2|0}(Q) = \text{Prob}_Q[Y^\dagger(2) = 1,\ Y^\dagger(0) = 0] \ , \tag{1.9}$$

which denotes the proportion of children who strictly benefit (in terms of the binary test score) from Head Start in comparison to home care. Note that this parameter can be re-written to satisfy Assumption 1.4.1 as follows

$$\text{PB}_{2|0}(Q) = \frac{\sum_{w \in \mathcal{W}_{2|0}} Q(w)}{\sum_{w \in \mathcal{W}} Q(w)} \ ,$$

where $\mathcal{W}_{2|0} = \{w \in \mathcal{W} : y(2) = 1, y(0) = 0\}$ is the set of all underlying values that correspond to a strictly higher outcome under Head Start in comparison to home school. As the converse of this parameter, I also consider

$$\text{PL}_{2|0}(Q) = \text{Prob}_Q[Y^\dagger(2) = 0,\ Y^\dagger(0) = 1] \ , \tag{1.10}$$

which denotes the proportion of children who strictly lose from Head Start in comparison to home care. Moreover, to simultaneously account for both these proportions, I also consider the difference in these quantities

$$\text{ATE}_{2|0}(Q) = \text{PB}_{2|0}(Q) - \text{PL}_{2|0}(Q) \ , \tag{1.11}$$

which corresponds to the average treatment effect (in terms of the binary test score) of Head Start versus home care. Note that in a manner similar to the re-writing of (1.9), the latter two parameters can also be re-written to satisfy Assumption 1.4.1. ■

Example 1.4.2. *(Option value of Head Start without an alternate preschool)* This class of parameters aims to evaluate the option value of Head Start when an alternate preschool option is absent, which corresponds to comparing policies allowing parents to freely select between Head Start and home care, i.e. from the choice set $\{2, 0\}$, versus mandating parents to select home care by not providing any preschool choices, i.e. through the choice set $\{0\}$. Note that, under this comparison, children whose parents do not exercise the Head Start option are left unaffected. To this end, denoting by

$$\mathcal{U}_{20|0} = \{u \in \mathcal{U} : d_{u,\{2,0\}} = 2\}$$

the set of preference types that prefer Head Start over home care, I consider

$$\text{PBOE}_{20|0}(Q) = \text{Prob}_Q[Y^\dagger(2) = 1,\ Y^\dagger(0) = 0 \mid U \in \mathcal{U}_{20|0}] \ , \tag{1.12}$$

which denotes the proportion who strictly benefit from the Head Start option conditional on the parents exercising the option in the setting where an alternate preschool option is

absent. Note that this parameter can be re-written to satisfy Assumption 1.4.1 as follows

$$\text{PBOE}_{20|0}(Q) = \frac{\sum\limits_{w \in \mathcal{W}_{20|0}} Q(w)}{\sum\limits_{w \in \mathcal{W}'_{20|0}} Q(w)} \,,$$

where $\mathcal{W}_{20|0} = \{w \in \mathcal{W} : y(0) = 0, y(2) = 1, u \in \mathcal{U}_{20|0}\}$ and $\mathcal{W}'_{20|0} = \{w \in \mathcal{W} : u \in \mathcal{U}_{20|0}\}$. Further, in order to evaluate the proportion of children whose parents exercise the option, I also consider

$$\text{PE}_{20|0}(Q) = \text{Prob}_Q[U \in \mathcal{U}_{20|0}] \tag{1.13}$$

which denotes the proportion of children whose parents prefer Head Start over home care. Note that in a manner similar to the rewriting of (1.9) and (1.12), this parameter can also be re-written to satisfy Assumption 1.4.1. ■

Example 1.4.3. *(Option value of Head Start with an alternate preschool)* The previous example focused on parameters that evaluated the option value of a Head Start preschool in the absence of an alternate preschool option. Here I list analogous parameters that evaluate this option value when an alternate preschool option is also available, i.e. the choice set $\{2, 1, 0\}$ versus the choice set $\{1, 0\}$. To this end, denoting by

$$\mathcal{U}_{210|10} = \{u \in \mathcal{U} : d_{u,\{2,1,0\}} = 2\}$$

the set of preference types that prefer Head Start over an alternate preschool and home care, let

$$\text{PBOE}_{210|10}(Q) = \text{Prob}_Q[Y^\dagger(2) = 1,\ Y^\dagger(D_{\{1,0\}}) = 0 \mid U \in \mathcal{U}_{210|10}] \tag{1.14}$$

denote the proportion who strictly benefit from the Head Start option conditional on the parents exercising the option in the setting where an alternate preschool option is available. Further, let

$$\text{PE}_{210|10}(Q) = \text{Prob}_Q[U \in \mathcal{U}_{210|10}] \tag{1.15}$$

denote the proportion of children whose parents prefer Head Start over an alternate preschool and home care. Similar to those in Example 1.4.2, note that these parameters can also be re-written to satisfy Assumption 1.4.1. ■

Remark 1.4.1. Previous studies on the HSIS have evaluated local average treatment effects in the framework of Imbens and Angrist (1994) - see Kline and Walters (2016) for arguments under which these parameters evaluate relevant policy effects. Using the notation in this paper, we could also consider parameters evaluated conditional on the so-called compliers such as, for example,

$$\text{PB}_{2|0,\text{late}}(Q) = \text{Prob}_Q[Y^\dagger(2) = 1,\ Y^\dagger(D(0)) = 0 \mid D(1) = 2,\ D(0) = d \in \{0,1\}] \ ,$$

which is a local counterpart of (1.9), where $D(z) = d_{U,C(z)}$ for $z \in \{0,1\}$. Similar to the conditional parameter in (1.12), such parameters can also be re-written to satisfy Assumption 1.4.1. As shown in Kline and Walters (2016), some of these local parameters can be

nonparametrically point identified under some conditions - see, for example, Angrist and Imbens (1995), Heckman and Pinto (2017), Heckman et al. (2006, 2008), Heckman and Vytlacil (2007), Hull (2015) and Kirkeboen et al. (2016) for further point identification results on such local parameters in multiple treatment settings. ■

Remark 1.4.2. A class of parameters that evaluate the spread of the distribution cannot be written as fractions of linear functions of the probability mass function Q as in Assumption 1.4.1. Examples of such parameters include the interquartile range and the variance. See Blundell et al. (2007) and Stoye (2010) for examples of studies that provide analytical bounds on such parameters in alternate settings. ■

1.4.2. Identification Problem

Given a pre-specified parameter of interest from the previous section, the identification problem studies what we can learn about it given the restrictions imposed on the unknown distribution of the model by the known distribution of the observed data and by the assumptions imposed on the model. Note that the identification problem abstracts away from sampling uncertainty, i.e. it supposes that the distribution of the observed data P is known rather than an estimate of it. The latter is addressed in Section 1.5.

I begin by describing the restrictions imposed by the distribution of the observed data P on the distribution of the model. These restrictions can formally be stated as

$$\sum_{w \in \mathcal{W}_x} Q(w) = P(x) \tag{1.16}$$

for all $x = (y, d, z, hc, cs) \in \mathcal{X}$, where $\mathcal{W}_x$ is the set of all w in $\mathcal{W}$ such that $c = c(1)z + c(1)(1-z)$, $d_{u,c} = d$ and $y = y(d)$. That is $\mathcal{W}_x$ is the set of all underlying values in $\mathcal{W}$

that could have possibly generated the observed value $x \in \mathcal{X}$ by the outcome equation in (1.3) and by the selection equation in (1.2).

Next, I describe the restrictions imposed by the experimental design of the HSIS, i.e. Assumption HSIS, on the distribution of the model. Before proceeding, I note that the developed identification analysis is flexible in the sense that it also allows for restrictions imposed by additional assumptions. More specifically, the identification analysis only requires that all these restrictions formally satisfy a specific condition. Similar to Assumption 1.4.1 and as discussed in Section 1.4.3, this condition is also motivated by those required to ensure that the final optimization problem is a linear program. In what follows, I first formally state this condition in the following assumption, and then show that the restrictions imposed by Assumption HSIS satisfy this condition.

Assumption 1.4.2. *Let $\mathcal{S}$ be a finite set of restrictions imposed on Q such that each restriction $s \in \mathcal{S}$ satisfies*

$$M_s(Q) \leq (or =)\ b_s\ , \tag{1.17}$$

where b_s is a known or identified value in $\mathbf{R}$, and

$$M_s(Q) = \sum_{w \in \mathcal{W}} a_s(w) \cdot Q(w)\ , \tag{1.18}$$

such that $a_s : \mathcal{W} \to \mathbf{R}$ is a known or identified function.

Assumption 1.4.2 states that there are only a finite number of restrictions imposed and that each restriction imposes a linear constraint on the distribution of the model. Moreover, the assumption allows these restrictions to be based on known values and on

values identified by features of the observed distribution of the data. As discussed in Section 1.5, this distinction is important when performing statistical inference.

In the following lemma, I show that Assumption HSIS imposes restrictions that satisfy the requirement in Assumption 1.4.2. Before proceeding, similar to Assumption 1.4.1, note that Assumption 1.4.2 is flexible in the sense that it also allows a researcher to impose additional assumptions on the model. In Section 1.6.1, I describe several additional nonparametric assumptions motivated by the details on the HSIS setting that satisfy this requirement.

Lemma 1.4.1. *Assumption HSIS imposes restrictions on Q that satisfy Assumption 1.4.2.*

PROOF: In order to see the restriction implied by Assumption HSIS(i), note that this assumption can be written as

$$\text{Prob}_Q[2 \notin C(1)] = 0 \ .$$

Equivalently, this can be re-written as a linear restriction on Q in the form of Assumption 1.4.2 as

$$\sum_{w \in \mathcal{W}_{\text{HSIS}}} Q(w) = 0 \ , \tag{1.19}$$

where $\mathcal{W}_{\text{HSIS}} = \{w \in \mathcal{W} : c(1) \in \{\{0\}, \{0,1\}\}\}$ is the set of all underlying values such that choice set with an offer does not contain a Head Start preschool.

In order to see the restrictions imposed by Assumption HSIS(ii), I first introduce some additional shorthand notation for the underlying values. To this end, denote by

$$\bar{y} = (y(0), y(1), y(2)) \in \{0,1\}^3$$

values of the potential outcomes and by

$$\bar{c} = (c(0), c(1)) \in \mathcal{C}^2$$

values of the potential choice sets, and by

$$\bar{z} = (hc, cs) \in \{0,1\}^2$$

values of the Head Start center characteristics. Using this notation, the conditional independence assumption in Assumption HSIS(ii) imposes the following restriction

$$\frac{Q(\bar{y}, u, \bar{c}, 0, \bar{z})}{\sum_{\bar{y}\in\{0,1\}^3, u\in\mathcal{U}, \bar{c}\in\mathcal{C}^2} Q(\bar{y}, u, \bar{c}, 0, \bar{z})} = \frac{Q(\bar{y}, u, \bar{c}, 1, \bar{z})}{\sum_{\bar{y}\in\{0,1\}^3, u\in\mathcal{U}, \bar{c}\in\mathcal{C}^2} Q(\bar{y}, u, \bar{c}, 1, \bar{z})}$$

for all values of $\bar{y} \in \{0,1\}^3$, $u \in \mathcal{U}$, $\bar{c} \in \mathcal{C}^2$ and $\bar{z} \in \{0,1\}^2$. Since the restrictions imposed by the distribution of the observed data imply that both the denominators can be written in terms of the known data distribution by

$$\sum_{\bar{y}\in\{0,1\}^3, u\in\mathcal{U}, \bar{c}\in\mathcal{C}^2} Q(\bar{y}, u, \bar{c}, z, \bar{z}) \quad = \sum_{y\in\{0,1\}, d\in\mathcal{D}} P(y, d, z, \bar{z})$$

for each $z \in \{0,1\}$, it then follows that above restriction can be re-written as a linear restriction on Q in the form of Assumption 1.4.2 as

(1.20)

$$\sum_{y\in\{0,1\},d\in\mathcal{D}} P(y,d,1,\bar{z})\cdot Q(\bar{y},u,\bar{c},0,\bar{z}) \quad - \sum_{y\in\{0,1\},d\in\mathcal{D}} P(y,d,0,\bar{z})\cdot Q(\bar{y},u,\bar{c},1,\bar{z}) = 0$$

for all values of $\bar{y} \in \{0,1\}^3$, $u \in \mathcal{U}$, $\bar{c} \in \mathcal{C}^2$ and $\bar{z} \in \{0,1\}^2$. ■

Given the restrictions imposed by the distribution of the observed data and by additional assumptions on the model distribution, I next ask what we can learn about the parameter of interest. This is formally defined by the identified set, i.e. the set of feasible parameter values such that the distribution of the model satisfies the various imposed restrictions. In order to formally state the identified set, denote first by $\mathbf{Q}_{\mathcal{W}}$ the set of all probability mass functions on the sample space $\mathcal{W}$. The identified set can then be stated as follows

(1.21) $$\Theta = \{\theta_0 \in \mathbf{R} : \theta(Q) = \theta_0 \text{ and } Q \in \mathbf{Q}\} \ ,$$

where

(1.22) $$\mathbf{Q} = \{Q \in \mathbf{Q}_{\mathcal{W}} : Q \text{ satisfies (1.16) and Assumption 1.4.2 } \}$$

is the set of all model distributions that satisfy the restriction imposed by the data and the assumptions. Note that if the identified set is empty then model is said to be mis-specified, i.e. the data is incompatible with the imposed assumptions on the model. Further, if the identified set contains a single point then the parameter is said to be point identified,

and if it is a proper subset of all possible values that the parameter can take then the parameter is set to be partially or set identified.

1.4.3. Identified Set

Due to the complicated structure of the underlying model, analytically characterizing the identified set in (1.21) is a difficult task. Instead, in the proposition below, I state an alternate more useful characterization of the identified set. This characterization in particular permits the use of tractable linear programming methods to obtain the identified set. Moreover, in the proof of this proposition presented below, I emphasize the specific role that conditions stated in Assumptions 1.4.1-1.4.2 play in ensuring that the optimization problem is a linear program. In this proposition below, I introduce the following additional quantity

$$\widetilde{\theta}(Q) = \sum_{w \in \mathcal{W}} a_{\text{num}}(w) \cdot Q(w) \ .$$

Proposition 1.4.1. *Suppose that* $\mathbf{Q}$ *in* (1.22) *is non-empty and the parameter of interest satisfies Assumption 1.4.1 such that*

$$\sum_{w \in \mathcal{W}} a_{den}(w) \cdot Q(w) > 0 \tag{1.23}$$

holds for every $Q \in \mathbf{Q}$*. Then the identified set in* (1.21) *can be written as*

$$\Theta = [\theta_l, \theta_u] \ , \tag{1.24}$$

where the lower and upper bounds of this interval are solutions to the following two linear programming problems

$$\theta_l = \min_{\gamma,\{Q(w)\}_{w\in\mathcal{W}}} \widetilde{\theta}(Q) \quad \textit{and} \quad \theta_u = \max_{\gamma,\{Q(w)\}_{w\in\mathcal{W}}} \widetilde{\theta}(Q) \ , \tag{1.25}$$

subject to the following constraints:

(i) $\gamma \geq 0$.

(ii) $0 \leq Q(w) \leq \gamma$ *for every* $w \in \mathcal{W}$.

(iii) $\sum_{w\in\mathcal{W}} Q(w) = \gamma$.

(iv) $\sum_{w\in\mathcal{W}_x} Q(w) = \gamma \cdot P(x)$ *for every* $x \in \mathcal{X}$.

(v) $\sum_{w\in\mathcal{W}} a_s(w) \cdot Q(w) \leq \gamma \cdot b_s$ *for every* $s \in \mathcal{S}$.

(vi) $\sum_{w\in\mathcal{W}} a_{den}(w) \cdot Q(w) = 1$.

PROOF: To begin, note that $\mathbf{Q}_{\mathcal{W}}$ is closed and convex. Further, note that $\mathbf{Q}$ is a set of distributions in $\mathbf{Q}_{\mathcal{W}}$ that is obtained by placing linear equality and inequality constraints imposed by the data in (1.16) and by restrictions characterized in Assumption 1.4.2. This in turn implies that $\mathbf{Q}$ is a closed and convex set as well.

Next, it follows from Assumption 1.4.1 that $\theta(Q)$ is a linear-fractional function of Q where as stated in the theorem the denominator is required to be positive, i.e. (1.23) holds, for every $Q \in \mathbf{Q}$. Along with $\mathbf{Q}$ being a closed and convex set, this in turn implies that $\theta(\mathbf{Q})$ is a closed and convex set in $\mathbf{R}$. More specifically, it follows that the identified set in (1.21) can be written as the closed interval in (1.24), where the lower bound and upper bound are given by

$$\theta_l = \min_{Q\in\mathbf{Q}} \theta(Q) \ \text{ and } \ \theta_u = \max_{Q\in\mathbf{Q}} \theta(Q) \ . \tag{1.26}$$

In order to complete the proof, note that both the optimization problems in (1.26) have linear-fractional objectives as guaranteed by Assumption 1.4.1 and a finite number of linear constraints as guaranteed by the data restrictions in (1.16) and by Assumption 1.4.2. Such optimization problem are commonly referred to as linear fractional programs. This structure required to ensure that the optimization problems in (1.26) can stated as linear-fractional programs is the fundamental value in considering Assumptions 1.4.1-1.4.2. This is useful because, as shown by Charnes and Cooper (1962), linear-fractional programs can be equivalently re-stated as linear programs of the form in (1.25) from which the result then directly follows.

To be specific, this restatement follows in two steps. First, it introduces the so-called Charnes-Cooper transformation

$$\widetilde{Q}(w) = \gamma \cdot Q(w) \quad \text{where} \quad \gamma = \frac{1}{\sum\limits_{w \in \mathcal{W}} a_{\text{den}}(w) \cdot Q(w)} \ ,$$

which is well defined given that (1.23) holds for every $Q \in \mathbf{Q}$. Then, by transforming the restrictions and parameters written in terms of Q to those in terms of γ and $\widetilde{Q}$, the minimization and maximization problem in (1.26) can be re-written as that in (1.25). ■

In the proof of Propositon 1.4.1 above, it is worth re-emphasizing that the main role that Assumptions 1.4.1-1.4.2 play is in ensuring that the optimization problems in (1.26) can be restated as linear programs. In principle, even in the absence of such linear assumptions, the identified set can still be characterized as solutions to optimization problems - see Torgovitsky (2016) for an example of such a result. Such optimization

problems may however not share the same reliability and tractability as linear programs for the dimensions present in this application, noted in Remark 1.4.5.

In addition to Assumptions 1.4.1-1.4.2, note that computing the identified set using Proposition 1.4.1 requires two further conditions. First, it requires $\mathbf{Q}$ to be non-empty, i.e. it requires the imposed model to be not mis-specified, to ensure that the identified set is an interval. If this is not the case, the linear program automatically terminates, which in turn provides a simple specification check for whether the assumptions imposed on the model are compatible with the observed data. Second, it requires the denominator of the parameter of interest to be positive, i.e. (1.23) holds, for every $Q \in \mathbf{Q}$. This is to ensure that the parameter of interest is well-defined for all feasible model distributions. This condition can easily be verified in practice. To see how, note that the denominator is also a linear-fractional parameter and more specifically a linear parameter as its corresponding denominator by construction is always one. The above proposition can then be employed to compute the lower bound for this auxiliary parameter to check if it is strictly positive given the restrictions imposed by the observed data and additional assumptions imposed on the model.

As noted when describing Assumptions 1.4.1-1.4.2, it is important to emphasize that an attractive feature of the linear programming method in Proposition 1.4.1 is its generality. In particular, this generality flexibly allows the study of various parameters under restrictions imposed by Assumption HSIS and also combinations of other assumptions. This permits in the empirical results in Section 1.6 to study various parameters for various specifications of restrictions without the need to analytically re-derive the identified set in each case.

The benefits of using computational procedures for obtaining the identified set has also been noted in previous studies on settings where deriving analytical bounds is difficult - see, for example, Manski (2007) in a related setting discussed in Remark 1.3.1 and Torgovitsky (2016) in a dynamic potential outcome setting. These studies have also proposed to focus attention on parameters and restrictions that are linear in the model distribution in order to use linear programs to compute the identified set. More specifically, with respect to the class of parameter, these previous proposals limit attention to only those that could be written as linear functions of the underlying model distribution. In contrast, by leveraging the result from Charnes and Cooper (1962), I also focus on a larger class of linear-fractional functions. In the setting of this paper, this in turn permits the study of an additional important class of conditional parameters such as the option value parameters described in Examples 1.4.2-1.4.3.

Remark 1.4.3. In absence of the binary transformations in (1.4) and (1.5), the model distribution is an infinite dimensional object as the test score outcome is continuously distributed. In this case, the identified set for similar parameters could also possibly be written as solutions to infinite dimensional programs under related conditions - see Mogstad et al. (2017) and Torgovitsky (2017) for examples of such results. However, computational methods that maximize, or minimize, over infinite dimensional spaces are in general intractable. To this end, the binary transformation considered here, and more generally any discrete transformation, translates the underlying space to a finite dimensional one, which allows for the use of tractable computational methods. ■

Remark 1.4.4. The identified set stated in (1.21) may not be sharp as only information from the distribution of the binary transformed test score is used rather than from underlying continuous test score. However, had the outcome of interest been binary, the identified set as stated would be sharp. Furthermore, had the outcome of interest been discrete, the analysis as previously noted can also be extended to obtain the sharpest set. ■

Remark 1.4.5. When employing Proposition 1.4.1 under the restrictions imposed by Assumption HSIS, there are 6144 unknown variables determined by the probability mass function of the model along with 12288 restrictions determined by (ii), 48 restrictions determined by (iv), and 3073 restrictions determined by (v). ■

1.5. Statistical Inference

The identification analysis in the previous section obtained the identified set assuming that the population distribution P of the observed data was known. The empirical results apply this analysis to obtain an estimate of the identified set using the empirical distribution of the HSIS sample data. Here, to be precise, the HSIS sample data is given by

$$X^{(n)} = \{X_g^{(n)} : 1 \leq g \leq G\} \tag{1.27}$$

where n denotes the total sample size, and

$$X_g^{(n)} = \{X_{ig} : 1 \leq i \leq n_g\}$$

denotes the cluster of observations of the random variable X in (1.7) for all the ith sampled children from the gth sampled Head Start center from the experiment.

In this section, I describe how I construct confidence intervals for the parameters of interest in the empirical results. Before proceeding, I note that deriving formal results for constructing confidence intervals for partially identified parameter that are uniformly asymptotically valid over a large class of distributions is outside the scope of this paper. As noted in Imbens and Manski (2004), ensuring uniform validity is in particular important in partially identified settings. In this section, I simply provide a description of how an existing method from the literature can practically be applied for the empirical results in order to sensibly capture the sampling uncertainty in the estimates.

I begin by noting that Kaido et al. (2016) and Mogstad et al. (2017) have recently proposed methods to obtain uniformly asymptotically valid confidence intervals for partially identified parameters obtained as solutions to linear programming problems. These methods specifically exploit the geometrical structure present in linear programming problems. Implementing these methods require solving non-linear optimization problems (over many bootstrap draws) that critically require that the obtained solution to be the global optimum rather than a local one. However, given the dimensions of the computational program noted in Remark 1.4.5 and the computational limitations noted in Remark 1.5.1, repeatedly solving such optimization problems reliably in this empirical application is not feasible.

Instead, I use a computationally tractable version of the profiled subsampling procedure proposed by Romano and Shaikh (2008a), which is shown to be uniformly asymptotically valid under some high level conditions - see also Politis and Romano (1994) and

Romano et al. (2012) for related results on the validity of subsampling. The sampling framework that the described method aims to capture is one where: (i) the random variables X_{ig} are identically distributed by P and uncorrelated across centers, i.e. independent across the index g; and (ii) the number of centers is large, i.e. $G \to \infty$, and the number of applicants per center is small, i.e. n_g is fixed for each g. The description of this method follows the exposition in Torgovitsky (2016), which restates the feasible set of underlying model distributions in (1.22) in terms of a moment equalities model.

To this end, first note that the restriction imposed by the data on the model distribution in (1.16) can be re-written as

$$E_P[m_{\text{dat},x}(X, Q)] = 0 \quad \text{for all } x \in \mathcal{X} \ , \tag{1.28}$$

where

$$m_{\text{dat},x}(X, Q) = \sum_{w \in \mathcal{W}_x} Q(w) - I\{X = x\} \ .$$

Next, suppose that the set of imposed restrictions $\mathcal{S}$ in Assumption 1.4.2 can be partitioned into $\mathcal{S}_1$ and $\mathcal{S}_2$, where each of the restrictions in these partitions satisfy some specific conditions stated in the following assumptions.

Assumption 1.5.1. *For all $s \in \mathcal{S}_1$, the linear restriction characterized by* (1.17) *is such that $a_s : \mathcal{W} \to \mathbf{R}$ and b_s are known and do not depend on P.*

Assumption 1.5.2. *For all $s \in \mathcal{S}_2$, the linear restriction characterized by* (1.17) *is an equality, and further is such that $a_s(w)$ for each $w \in \mathcal{W}$ and b_s are linear functions of P.*

Assumption 1.5.1 states that all restrictions imposed by $s \in \mathcal{S}_1$ are deterministic in the sense that they do not depend on features of the data and are hence not estimated. The restriction imposed by Assumption HSIS(i) in (1.19), for example, satisfies this assumption. The restrictions satisfying this assumption determine the feasible "parameter" space of the model, i.e.

$$\mathbf{Q}' = \{Q \in \mathbf{Q}_{\mathcal{W}} : M_s(Q) \leq b_s \text{ for s} \in \mathcal{S}_1\} \ . \tag{1.29}$$

In contrast, Assumption 1.5.2 states all restrictions imposed by $s \in \mathcal{S}_2$ are stochastic in the sense that they depend linearly on the distribution of the data and are hence estimated using the sample data. The restrictions imposed by Assumption HSIS(ii) in (1.20), for example, satisfy this assumption. Moreover, this assumption implies that each restriction (1.17) for $s \in \mathcal{S}_2$ can be re-written as a moment equality, i.e.

$$E_P[m_s(X, Q)] = 0 \ \text{ for all } s \in \mathcal{S}_2 \ , \tag{1.30}$$

which can be formally shown using the same arguments as used for showing the moment equalities imposed by the data restrictions. Using this notation, we can see that the feasible set of underlying model distributions is defined by the parameter space in (1.29) and by the moment equalities in (1.28) and (1.30).

Using the restated moment equalities model, I now describe a test at a pre-specified level $\alpha \in (0, 1)$ based on Romano and Shaikh (2008a) for a null hypothesis that the parameter of interest is equal to a given value $\theta_0 \in \mathbf{R}$, i.e. $\theta(Q) = \theta_0$ - see Canay and Shaikh (2017) for a review of other methods that can be applied in moment equalities models. Confidence intervals can then be obtained by inverting the test, i.e. by collecting

the set of all values of θ_0 for which the test does not reject at level α. The test requires a choice of test statistic that rejects the null hypothesis for large values. I propose using the following profiled test statistic

$$TS_n(\theta_0) = \sqrt{G} \min_{Q \in \mathbf{Q}(\theta_0)} \sum_{x \in \mathcal{X}} |\hat{m}_{\text{dat},x}(Q)| + \sum_{s \in \mathcal{S}_1} |\hat{m}_s(Q)| \ , \tag{1.31}$$

where $\hat{m}_{\text{dat},x}(Q)$ is an empirical analogue of the the moment in (1.28), $\hat{m}_s(Q)$ is an empirical analogue of the moment in (1.30), and $\mathbf{Q}(\theta_0) = \{Q \in \mathbf{Q}' : \theta(Q) = \theta_0\}$. As shown in Appendix A.0.1, an important practical benefit of this choice of test statistic is that it can be solved as a solution to a linear programming problem. In order to obtain a critical value, the test draws B subsamples of size b drawn randomly without replacement from the clusters of Head Start center observations in (1.27) and computes $TS_{n,b,j}(\theta_0)$, which is the test statistic in (1.31) using the jth subsample of size b. Denoting by $\hat{c}_n(1-\alpha, \theta_0)$ the $(1-\alpha)$-quantile of the subsampling distribution

$$L_n(t, \theta_0) = \frac{1}{B} \sum_{j=1}^{B} 1\{TS_{n,b,j}(\theta_0) \leq t\} \ ,$$

the test can be then be denoted by

$$\phi_n^{SS}(\theta_0) = 1\{TS_n(\theta_0) \geq \hat{c}_n(1-\alpha, \theta_0)\} \ .$$

The $(1-\alpha)$ confidence interval for the parameter of interest is in turn given by

$$C_n = \{\theta_0 \in \mathbf{R} : \phi_n^{SS}(\theta_0) = 0\} \ .$$

Note that the above described test requires computing the test statistic over multiple subsamples and further requires inverting the test to construct confidence intervals. Despite this, by ensuring that all the performed optimization problems are linear programs, I find that constructing confidence intervals using this test can still be relatively tractable.

Remark 1.5.1. Constructing confidence intervals can be an expensive computational exercise due to the fact that the test statistic needs to be computed over many bootstrap draws or subsamples and further due to the inversion of the test. For the purposes of the HSIS data, due to a restricted access data agreement, an important computational limitation was that this exercise needed to be performed on a personal computer without any assistance from a computing cluster. ■

Remark 1.5.2. The subsampling test requires a choice of block size b. Formally, the only requirements are that $b \to \infty$ and $\frac{b}{G} \to 0$ as $G \to \infty$. In the empirical results, I take $b = G^{2/3}$ following Bugni (2016). Moreover, I take the number of drawn subsamples B to be 100 due to the computational limitations noted in Remark 1.5.1. ■

Remark 1.5.3. By drawing subsamples from the clusters of Head Start centers, the described test accounts for the dependence present between observations sampled from a given Head Start center. Randomized experiments may however introduce additional dependence due to various randomization schemes used to assign treatment that the described test does not capture. See Bugni et al. (2017a,b) for formal results on performing valid inference for point identified parameters under such dependence introduced by a general class of randomization schemes. ■

1.6. Empirical Results

In this section, I present the empirical results using the formal analysis developed in the preceding sections on the data from the HSIS. Table 1.2 begins by presenting the estimated identified sets and 95% confidence intervals for the parameters discussed in Section 1.4.1 using only the restrictions imposed by the observed data and Assumption HSIS.

Observe in Table 1.2 that the lower bound of the parameter $\text{PE}_{210|10}$, the proportion of children whose parents exercise the Head Start option when an alternate preschool is available, is equal to zero. Since this parameter corresponds to the denominator of the parameter $\text{PBOE}_{210|10}$ in Example 1.4.3, the identified set for the latter parameter hence cannot be computed. In particular, the linear programming method in Proposition 1.4.1 cannot be employed since, as noted in the previous section, it requires that the denominator for the parameter of interest is strictly positive under the feasible model distributions. In order to still measure the benefits of a Head Start option when an alternative preschool is present, using the notation from Section 1.4.1, I additionally report results for

$$\text{PB}_{210|10}(Q) = \text{Prob}_Q[Y^\dagger(D_{\{2,1,0\}}) = 1,\ Y^\dagger(D_{\{1,0\}}) = 0]\ , \tag{1.32}$$

which denotes the proportion of children who strictly benefit from a Head Start option when an alternate preschool option is also present. However, as further noted in Section 1.6.2, the lower bound of the denominator can be strictly positive under restrictions imposed by additional assumptions, which then permits the study of $\text{PBOE}_{210|10}$ under these assumptions.

Table 1.2. Estimated worst case bounds and confidence intervals for each age group.

Parameter	Age 3	Age 4
	Head Start versus Home care	
$PB_{2\|0}$	[0.000, 0.537]	[0.000, 0.627]
	[0.000, 0.689]	*[0.000, 0.693]*
$PL_{2\|0}$	[0.000, 0.307]	[0.000, 0.432]
	[0.000, 0.460]	*[0.000, 0.572]*
$ATE_{2\|0}$	[-0.149, 0.386]	[-0.280, 0.420]
	[-0.239, 0.506]	*[-0.400, 0.578]*
	Option value of Head Start without an alternate preschool	
$PBOE_{20\|0}$	[0.000, 0.594]	[0.000, 0.696]
	[0.000, 0.694]	*[0.000, 0.796]*
$PE_{20\|0}$	[0.866, 0.917]	[0.799, 0.911]
	[0.786, 0.959]	*[0.729, 0.960]*
	Option value of Head Start with an alternate preschool	
$PB_{210\|10}$	[0.000, 0.670]	[0.000, 0.532]
	[0.000, 0.730]	*[0.000, 0.622]*
$PBOE_{210\|10}$	-	-
	-	-
$PE_{210\|10}$	[0.000, 0.866]	[0.000, 0.799]
	[0.000, 0.930]	*[0.000, 0.893]*

Notes: For each parameter, the upper panel denotes the estimated identified set and the lower panel denotes the 95% confidence interval. Appendix A.0.3 provides details on how the data used for these results was constructed.

The so-called worst case bounds in Table 1.2 are substantially informative on inferring how parents choose between the a Head Start preschool and home care when a Head Start offer is made, i.e. the parameter denoted by $PE_{20|0}$. Using only the data and the experimental design of the HSIS, the estimated identified sets implies that between 79.9% to 91.7% of the parents across both age cohorts exercise their Head Start option when only home care is the outside option. However, the data and the experimental design are generally uninformative with respect to reaching stronger conclusions based on the remaining parameters. For example, one cannot infer whether Head Start positively or negatively affects test scores as measured across various parameters as zero is included in the identified set for these parameters.

In the remainder of this section, I illustrate additional identifying assumptions with the aim of tightening the worst case bounds and reaching stronger conclusions regarding the efficacy of Head Start. In order to easily employ the linear programming method in Proposition 1.4.1, note that these assumptions impose restrictions that satisfy Assumption 1.4.2. In Section 1.6.1 below, I first describe these assumptions and provide a discussion on their plausibility in the context of the HSIS. In Section 1.6.2 below, I then report the estimated bounds under these assumptions and provide a discussion on their identifying power.

1.6.1. Additional Identifying Assumptions

An important feature of the below described nonparametric assumptions is their transparency in terms of the restrictions that they impose on the empirical setting. This allows us to easily assess whether these restrictions are justified in the context of the HSIS - see Manski (2003) for a general discussion on the importance of such an assessment. In the description of these assumptions below, I focus attention to this assessment and leave the explicit mathematical derivations of their imposed restrictions satisfying Assumption 1.4.2 for Appendix A.0.2. Furthermore, from the explicit forms of these restrictions, it is apparent that they satisfy either Assumption 1.5.1 or Assumption 1.5.2, which in turn allows constructing confidence intervals using the subsampling procedure discussed in Section 1.5.

1.6.1.1. Unaltered Alternatives. Recall that Assumption HSIS required that the experimental offer guaranteed that a Head Start preschool was in the choice set. The assumption however imposed no restrictions on the choice set had an experimental offer

not been made. In the following assumption, I state a reasonable assumption on the relation between the choice set with and without an offer.

Assumption UA. $C(1) = C(0) \cup \{2\}$.

Assumption UA states that apart from introducing a Head Start preschool in the child's choice set of care setting, it leaves the choice set completely unaltered. This assumption makes an implicit restriction on the application decisions of the parents to obtain preschool slots, i.e. the unobserved process that generates the choice sets. For example, it rules out cases such as not obtaining an experimental offer induced the parents to apply to additional alternate preschools and successfully obtain an offer, i.e. a case where there is an alternate preschool present in $C(0)$ but not in $C(1)$.

1.6.1.2. Site Level Instruments. Previous studies on the HighScope curriculum and on class size ratios suggest that the considered Head Start preschool characteristics would affect the child's test scores under Head Start. Moreover, due to competition between preschools, it is likely that they would also affect the child's test scores under an alternative preschool. However, it is reasonable to believe that it would play no role on the child's test score under home care, which motivates the following assumption.

Assumption SLI. $Y(0) \perp (HC, CS)$.

Assumption SLI states that the considered baseline covariates of a Head Start preschool are independent of the child's test scores had the child attended home care. Such an assumption rules out settings where the characteristics of a Head Start preschool are endogenously determined by the quality of families applying there, and hence the outcome under home care.

Remark 1.6.1. Kline and Walters (2016) use Head Start preschool characteristics with additional assumptions to point identify local average treatment effects on subpopulations other than the compliers, described in Remark 1.4.1. These additional assumptions essentially require that the local average treatment effects conditional on values of the Head Start characteristics are homogeneous across these values. ■

Remark 1.6.2. It may also be reasonable to assume that the test scores under Head Start weakly improve with the described Head Start characteristics in the sense of the monotone instrumental variable assumption of Manski and Pepper (2000, 2009). I find that this assumption does not have much identifying power in this setting. Note however that such an assumption may have more identifying power in alternate applications. ■

1.6.1.3. Monotone Treatment Response. A large body of research on early childhood interventions suggest that preschool enrollment may be beneficial in comparison to home care. In turn, it may be reasonable to believe that both Head Start and alternate preschools are at least as good with respect to a child's test score in comparison to home care, which motivates the following assumption.

Assumption MTR. *For each* $d \in \{1, 2\}$, $Y(d) \geq Y(0)$.

Assumption MTR states that the potential test scores under both preschools is weakly greater than the potential test score under home care. This assumption is a version of the Monotone Treatment Response (MTR) assumption proposed by Manski (1997b). In particular, note that such an assumption imposes that no child can possibly lose from attending any of the two preschools in comparison to home care. To the extent that some

children may strictly be better off under home care, such an assumption may be subject to suspect.

1.6.1.4. Roy Model. The parents' preferences for care settings may be related to the potential test score their child may receive under each of these settings. In the case where it is assumed that these preferences are based solely on optimizing these potential test scores, we obtain a version of the Roy model stated as follows.

Assumption Roy. *For each* $d, d' \in \mathcal{D}$*, if* $Y(d') > Y(d)$ *then* $d_{U,\{d,d'\}} = d'$.

Assumption Roy states that parents are omniscient in the sense that they exactly know their child's potential test scores under each setting and then prefer the care setting where their child would attain the highest test score. To be more clear, note that the above assumption can equivalently be restated in terms of utilities as

$$Y(d') > Y(d) \implies U(d') > U(d)$$

for every $d, d' \in \mathcal{D}$, i.e. if the potential test score under a given treatment is higher than that under another then the utility under that treatment is also higher than the utility under the other. The credibility of this assumption can be suspect for multiple reasons of which I state two important ones relevant to this setting. First, parents may not exactly know their child's potential test score under each care setting but may only have an expectation of what it may be. In turn, parents' preferences may be based on selecting care settings that maximize expected rather than actual potential test scores. To the extent that these two values differ, the Roy model may not reasonable capture how preferences may be related to the potential test scores. Second, parents may make

enrollment decisions based not only on test scores but also on additional factors such as possible costs that may be incurred from attending a given preschool.

Assumption Roy can be weakened to partially address the second concern noted above. Since the preschools may possible be closely related, it is reasonable to believe that costs incurred at either are similar. Hence, it may be the case such that at least the preferences between the two preschools are based solely on potential test scores as stated in the following assumption.

Assumption SemiRoy. *For each* $d, d' \in \{1, 2\}$, *if* $Y(d') > Y(d)$ *then* $d_{U,\{d,d'\}} = d'$.

Remark 1.6.3. Assumption Roy has an important distinction from versions of the Roy model previously studied in the literature such as, for example, in Heckman and Honore (1990) and Mourifie et al. (2015). These versions do not account for the possibility that the treatment is not selected from the set of all treatments, i.e. it does not account for unobserved heterogeneity in choice sets. ■

1.6.2. Bounds under Additional Identifying Assumptions

Table 1.3 and Table 1.4 report for the age 3 cohort and the age 4 cohort, respectively, the estimated identified set for the parameters presented in Table 1.2 under the additional restrictions imposed by the various assumptions described in the previous section. I organize the discussion of the results by the various assumptions:

UA: Assumption UA has identifying power in terms of the proportion who benefit, and proportion who are induced to enroll into Head Start when provided the option in the presence of an alternate preschool option. It provides evidence

Table 1.3. Estimated identified sets for the age 3 cohort

Assumption									
UA	✓				✓	✓	✓	✓	✓
SLI						✓	✓		✓
MTR		✓			✓	✓	✓		
Roy			✓					✓	✓
SemiRoy				✓			✓		
Parameter									
	Head Start versus Home care								
$PB_{2\|0}$	0.009	0.029	0.029	0.000	0.070	0.096	0.096	0.070	0.096
	0.537	0.386	0.352	0.537	0.386	0.320	0.309	0.352	0.286
$PL_{2\|0}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.307	-0.000	0.100	0.307	0.000	0.000	0.000	0.100	0.100
$ATE_{2\|0}$	-0.149	0.029	-0.031	-0.149	0.070	0.096	0.096	-0.031	0.027
	0.386	0.386	0.352	0.375	0.386	0.320	0.309	0.352	0.286
	Option value of Head Start without an alternate preschool								
$PBOE_{20\|0}$	0.010	0.032	0.032	0.000	0.076	0.105	0.105	0.076	0.105
	0.561	0.414	0.407	0.594	0.396	0.335	0.326	0.388	0.326
$PE_{20\|0}$	0.866	0.866	0.866	0.866	0.866	0.866	0.866	0.866	0.866
	0.917	0.917	0.917	0.917	0.917	0.917	0.917	0.917	0.917
	Option value of Head Start with an alternate preschool								
$PB_{210\|10}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	0.664	0.340	0.340	0.670	0.305	0.275	0.167	0.167	0.167
$PBOE_{210\|10}$	0.000	-	-	-	0.000	0.000	0.000	0.000	0.000
	0.822	-	-	-	0.679	0.657	0.533	0.533	0.533
$PE_{20\|0}$	0.207	0.000	0.000	0.000	0.207	0.207	0.207	0.207	0.207
	0.866	0.866	0.866	0.866	0.866	0.866	0.866	0.866	0.866

Notes: For each parameter and combination of assumptions denoted by the various check marks, the upper panel denotes the lower bound of the identified set and the lower panel denotes the upper bound of the identified set. Appendix A.0.3 provides details on how the data used for these results was constructed.

that there is a strictly positive proportion of three-year-old children, at least approximately 1%, who benefit from Head Start in comparison to home care. It is also provides evidence that a strictly positive proportion of parents from 20.7% to 86.6% across the two age cohorts are induced to enroll their child into Head Start when provided the option when an alternate preschool option was available to them.

SLI: In unreported results, I find that Assumption SLI does not have much identifying power in this setting in terms of reaching stronger conclusions over what

Table 1.4. Estimated identified sets for the age 4 cohort

Assumption									
UA	✓				✓	✓		✓	
SLI						✓	✓		✓
MTR		✓			✓	✓	✓		
Roy			✓					✓	✓
SemiRoy				✓			✓		
Parameter									
	Head Start versus Home care								
$PB_{2\|0}$	0.000	0.000	0.000	0.000	0.045	0.045	0.000	∅	0.000
	0.627	0.420	0.365	0.624	0.420	0.376	0.351		0.321
$PL_{2\|0}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	∅	0.000
	0.432	-0.000	0.146	0.432	0.000	0.000	0.000		0.146
$ATE_{2\|0}$	-0.280	0.000	-0.101	-0.280	0.045	0.045	0.000	∅	-0.089
	0.420	0.420	0.365	0.395	0.420	0.376	0.351		0.321
	Option value of Head Start without an alternate preschool								
$PBOE_{20\|0}$	0.000	0.000	0.000	0.000	0.049	0.049	0.000	∅	0.000
	0.654	0.473	0.456	0.695	0.427	0.399	0.402		0.402
$PE_{20\|0}$	0.799	0.799	0.799	0.799	0.799	0.799	0.799	∅	0.799
	0.911	0.911	0.911	0.911	0.911	0.911	0.911		0.911
	Option value of Head Start with an alternate preschool								
$PB_{210\|10}$	0.000	0.000	0.000	0.000	0.000	0.000	0.000	∅	0.000
	0.523	0.333	0.333	0.532	0.262	0.262	0.310		0.310
$PBOE_{210\|10}$	0.000	-	-	-	0.000	0.000	-	∅	-
	0.738	-	-	-	0.585	0.585	-		-
$PE_{20\|0}$	0.276	0.000	0.000	0.000	0.276	0.276	0.000	∅	0.000
	0.799	0.799	0.797	0.799	0.799	0.799	0.797		0.797

Notes: For each parameter and combination of assumptions denoted by the various check marks, the upper panel denotes the lower bound of the identified set and the lower panel denotes the upper bound of the identified set. Appendix A.0.3 provides details on how the data used for these results was constructed.

can already be learned from the worst-case bounds. However, when combined with other assumptions, they can have identifying power as discussed below.

MTR: Assumption MTR provides evidence that there is a strictly positive proportion of three-year-old children, from 2.9% to 38.6% and from 3.2% to 41.4%, who benefit from Head Start as measured across two alternate policy dimensions by $PB_{2|0}$ and $PBOE_{20|0}$, respectively. Moreover, since by construction this assumption imposes that there is a zero proportion of children who lose from Head

Start, it implies that $PB_{2|0}$ is equal to $ATE_{2|0}$, the average treatment effect of Head Start in comparison to home care, which is then also strictly positive.

Roy + SemiRoy: Similar to Assumption MTR, Assumption Roy also provides evidence that there is a strictly positive proportion of three-year-old children, from 2.9% to 35.2% and from 3.2% to 40.7%, who benefit from Head Start as measured by $PB_{2|0}$ and $PBOE_{20|0}$, respectively. However, in contrast, it allows for at most 10% of children who are strictly better off under home care as observed in $PL_{2|0}$. Comparing the results between Assumption Roy and Assumption SemiRoy implies that using the total preference ordering between the care settings is important in this setting in terms of identifying power.

Combinations: Combining Assumptions UA-MTR strengthens the estimated bounds for the three-year-old children from 7.0% to 38.6% and from 7.6% to 39.6% for the parameters $PB_{2|0}$ and $PBOE_{20|0}$, respectively. This combination additionally provides evidence that there is strictly positive proportion of four-year-old children, from 4.5% to 42.0% and from 4.9% to 42.7%, who benefit from Head Start as measured by $PB_{2|0}$ and $PBOE_{20|0}$, respectively. This additional evidence however is absent when the assumptions are specified separately. These estimated identified sets are further tightened when Assumptions SLI-SemiRoy are also sequentially imposed. When combining these assumptions with Assumption Roy instead of Assumption MTR, similar results are found for the age three cohort. However, for the age four cohort, a combination of Assumption UA, and Assumption Roy or Assumption SemiRoy is considered infeasible by the linear program, i.e. the data is not compatible with these combinations of assumptions.

As noted above, the estimated identified sets across various specifications of assumptions are fairly informative for the majority of the parameters. An exception is the estimated identified set for the parameters that measure the proportion of children who benefit from the option value of Head Start when an alternative preschool is available, i.e. $PB_{210|10}$ and $PBOE_{210|10}$. This is primarily due to the fact that no assumptions are imposed on whether an alternative preschool is present in a child's choice set by the experimental setting or by the additional identifying assumptions.

Overall, across various specifications of assumptions, the evidence suggested by the estimated identified sets can be qualitatively summarized as follows: (i) Head Start is more effective in comparison to home school in terms of improving short-term test scores; (ii) providing a Head Start option is effective in terms of inducing people to attend Head Start and further benefiting those children who attend in terms of short-term test scores; (iii) these summarized results are more pronounced for the age three cohort than the age four cohort; and (iv) results on the whether there is a strictly positive effect of providing a Head Start option when an alternate is present is inconclusive, though an upper value that bounds the possible effect is fairly tight.

Recall from Section 1.5 that due to the practical limitations computing confidence intervals for all the combinations of parameters and specifications of assumptions reported above is difficult. For the purposes of illustration, I only report and discuss 95% confidence intervals for some choices of parameters and assumptions. For the parameters $PB_{2|0}$ and $PBOE_{20|0}$, I find that: (i) under the combination of Assumption UA and Assumption MTR, the confidence interval is $[0.000, 0.506]$ and $[0.000, 0.526]$, respectively; whereas (ii) under the combination of Assumption UA, Assumption MTR and Assumption SLI, the

confidence interval is $[0.000, 0.570]$ and $[0.000, 0.505]$, respectively. In summary, at least for these few choice of parameters and assumptions, I find that the positive benefits of the Head Start using estimated identified sets are not statistically significant using the subsampling procedure at a 95% confidence level. However, as previously observed for example by Torgovitsky (2016) in a related setting, it is possible that the subsampling confidence intervals can be highly conservative in this setting. I leave the use of tractable, less conservative and valid statistical inference methods in this empirical application for future work.

1.7. Conclusion

This paper studies the identification of program effects in settings with latent choice sets. This analysis was developed in the context of the Head Start Impact Study (HSIS) social experiment. In particular, the analysis required a deep understanding of the experimental design and the setting in order to leverage useful information to reach informative conclusions on various parameters of interests.

More specifically, to explicitly account for the latent choice sets, I developed a nonparametric model that was tightly connected to the experimental design of the study. For a large class of parameters in this model, I showed that the identified set using the information provided by the experimental design and various unique institutional details of the setting can be constructed using a flexible linear programming method. Applying the developed analysis, I found that Head Start is effective across various policy dimensions, which corroborates the positive findings of previous studies under weak nonparametric assumptions.

To conclude, I would like to emphasize that though developed in the context of the HSIS the underlying framework applies more generally. In particular, it directly extends to alternate experimental settings with latent choice sets and can also be extended to observational settings with latent choice sets. Indeed, the described parameters of interest and identifying assumptions can be viewed as as illustrative examples with similar counterparts in the alternate empirical settings.

CHAPTER 2

Approximate Permutation Tests and Induced Order Statistics in the Regression Discontinuity Design

2.1. Introduction

The regression discontinuity design (RDD) has been extensively used in recent years to retrieve causal treatment effects - see Lee and Lemieux (2010a) and Imbens and Lemieux (2008a) for exhaustive surveys. The design is distinguished by its unique treatment assignment rule. Here individuals receive treatment when an observed covariate (commonly referred to as the running variable) crosses a known cutoff or threshold, and receive control otherwise. Hahn et al. (2001a) illustrates that such an assignment rule allows nonparametric identification of the average treatment effect (ATE) at the cutoff, provided that potential outcomes have continuous conditional expectations at the cutoff. The credibility of this identification strategy along with the abundance of such discontinuous rules in practice have made the RDD increasingly popular in empirical applications.

The continuity assumption that is necessary for nonparametric identification of the ATE at the cutoff is fundamentally untestable. Empirical studies however assess the plausibility of their RDD by exploiting two testable implications of a stronger identification assumption proposed by Lee (2008). We can describe the two implications as follows: (i) individuals have imprecise control over the running variable, which translates into the density of the running variable being continuous at the cutoff; and (ii) the treatment is

locally randomized at the cutoff, which translates into the distribution of all observed baseline covariates being continuous at the cutoff. The second prediction is particularly intuitive and, quite importantly, analogous to the type of restrictions researchers often inspect or test in a fully randomized controlled experiment. The practice of judging the reliability of RDD applications by assessing either of the two above stated implications (commonly referred to as manipulation, falsification or placebo tests) is ubiquitous in the empirical literature.[1] However, in regards to the second testable implication, researchers often verify continuity of the means of baseline covariates at the cutoff, which is a weaker requirement than Lee's implication.

This paper proposes a novel permutation test for the null hypothesis on the second testable implication, i.e., the distribution of baseline covariates is continuous at the cutoff.[2] The new test has a number of attractive properties. First, our test controls the limiting null rejection probability under fairly mild conditions, and delivers finite sample validity under stronger, but yet plausible, conditions. Second, our test is more powerful against some alternatives than those aimed at testing the continuity of the *means* of baseline covariates at the cutoff, which appears to be a dominant practice in the empirical literature. Third, our test is arguably simple to implement as it only involves computing order statistics and empirical cdfs with a fixed number of observations closest to the cutoff.

[1]Table B.1 surveys RDD empirical papers in four leading applied economic journals during the period 2011-2015, see Appendix B.0.5 for further details. Out of 62 papers, 43 of them include some form of manipulation, falsification or placebo test. In fact, the most popular practice involves evaluating the continuity of the *means* of baseline covariates at the cutoff (42 papers).

[2]It is important to emphasize that the null hypothesis we test in this paper is neither necessary nor sufficient for identification of the ATE at the cutoff. See Section 2 for a discussion on this.

This contrasts with the few existing alternatives that require local linear estimation, undersmoothing, and delicate bandwidth choices. Finally, we have developed a companion Stata package to facilitate the adoption of our test.[3]

The construction of our test is based on the simple intuition that observations close to the cutoff are *approximately* (but not exactly) identically distributed to either side of it when the null hypothesis holds. This allows us to permute these observations to construct an *approximately* valid test. In other words, the formal justification for the validity of our test is asymptotic in nature and recognizes that traditional arguments advocating the use of permutation tests are not necessarily valid under the null hypothesis of interest; see Section 2.3.2 for a discussion on this distinction. The novel asymptotic framework we propose aims at capturing a *small* sample problem: the number of observations close to the cutoff is often small even if the total sample size is large. More precisely, our asymptotic framework is one in which the number of observations q that the test statistic contains from either side of the cutoff is *fixed* as the total sample size n goes to infinity. Formally, we exploit the recent asymptotic framework developed by Canay, Romano and Shaikh (2017) for randomization tests, although we introduce novel modifications to accommodate the RDD setting. Further, in an important intermediate stage, we use induced order statistics, see Bhattacharya (1974) and (2.8), to frame our problem and develop some insightful results of independent interest in Theorem 2.4.1.

An important contribution of this paper is to show that permutation tests can be justified in RDD settings through a novel asymptotic framework that aims at embedding a small sample problem. The asymptotic results are what primarily separates this paper

[3]The Stata package rdperm can be downloaded from `http://sites.northwestern.edu/iac879/software/`.

from others in the RDD literature that have advocated for the use of permutation tests (see, e.g., Cattaneo et al., 2015; Sales and Hansen, 2015; Ganong and Jäger, 2015). In particular, all previous papers have noticed that permutation tests become appropriate for testing null hypotheses under which there is a *neighborhood* around the cutoff where the RDD can be viewed as a randomized experiment. This, however, deviates from traditional RDD arguments that require such local randomization to hold only *at the cutoff*. Therefore, as explained further in Section 2.3.2, this paper is the first to develop and to provide formal results that justify the use of permutation tests asymptotically for these latter null hypotheses. Another contribution of this paper is to exploit the testable implication derived by Lee (2008), which is precisely a statement on the distribution of baseline covariates, and note that permutation tests arise as natural candidates to consider. Previous papers have focused attention on hypotheses about distributional treatment effects, which deviates from the predominant interest in ATEs, and do not directly address the testing problem we consider in this paper.

The remainder of the paper is organized as follows. Section 2.2 introduces the notation and discusses the hypothesis of interest. Section 2.3 introduces our permutation test based on a fixed number of observations closest to the cutoff, discusses all aspects related to its implementation in practice, and compares it with permutation tests previously proposed in the RDD setting. Section 2.4 contains all formal results, including the description of the asymptotic framework, the assumptions, and the main theorems. Section 2.5 studies the finite sample properties of our test via Monte Carlo simulations. Finally, Section 2.6 implements our test to reevaluate the validity of the design in Lee (2008), a familiar application of the RDD to study incumbency advantage. All proofs are in the Appendix.

2.2. Testable implications of local randomization

Let $Y \in \mathbf{R}$ denote the (observed) outcome of interest for an individual or unit in the population, $A \in \{0, 1\}$ denote an indicator for whether the unit is treated or not, and $W \in \mathcal{W}$ denote observed, baseline covariates. Further denote by $Y(1)$ the potential outcome of the unit if treated and by $Y(0)$ the potential outcome if not treated. As usual, the (observed) outcome and potential outcomes are related to treatment assignment by the relationship

$$Y = Y(1)A + Y(0)(1 - A) \ . \tag{2.1}$$

The treatment assignment in the (sharp) Regression Discontinuity Design (RDD) follows a discontinuous rule,

$$A = I\{Z \geq \bar{z}\} \ ,$$

where $Z \in \mathcal{Z}$ is a scalar random variable known as the running variable and $\bar{z}$ is the threshold or cutoff value. For convenience, we normalize $\bar{z} = 0$. This treatment assignment rule allows us to identify the average treatment effect (ATE) at the cutoff; i.e.,

$$E[Y(1) - Y(0)|Z = 0] \ .$$

In particular, Hahn et al. (2001a) establish that identification of the ATE at the cutoff relies on the discontinuous treatment assignment rule and the assumption that

$$E[Y(1)|Z = z] \quad \text{and} \quad E[Y(0)|Z = z] \quad \text{are both continuous in } z \text{ at } z = 0 \ . \tag{2.2}$$

Reliability of the RDD thus depends on whether the mean outcome for units marginally below the cutoff identifies the true counterfactual for those marginally above the cutoff.

Despite the continuity assumption appearing weak, Lee (2008) states two practical limitations for empirical researchers. First, it is difficult to determine whether the assumption is plausible as it is not a description of a treatment-assigning process. Second, the assumption is fundamentally untestable. Motivated by these limitations, Lee (2008, Condition 2b) considers an alternative (and arguably stronger) sufficient condition for identification. The new condition is intuitive and leads to clean testable implications that are easy to assess in an applied setting. In RDD empirical studies, these implications are often presented (with different levels of formality) as falsification, manipulation, or placebo tests (see Table B.1 for a survey).

In order to describe Lee's alternative condition, let U be a scalar random variable capturing the unobserved type or heterogeneity of a unit in the population. Assume there exist measurable functions $m_0(\cdot)$, $m_1(\cdot)$, and $m_w(\cdot)$, such that

$$Y(1) = m_1(U), \quad Y(0) = m_0(U), \quad \text{and} \quad W = m_w(U) \ .$$

Condition 2b in Lee (2008) can be stated in our notation as follows.

Assumption 2.2.1. *The cdf of Z conditional on U, $F(z|u)$, is such that $0 < F(0|u) < 1$, and is continuously differentiable in z at $z = 0$ for each u in the support of U. The marginal density of Z, $f(z)$, satisfies $f(0) > 0$.*

This assumption has a clear behavioral interpretation - see Lee (2008) and Lee and Lemieux (2010a) for a lengthly discussion of this assumption and its implications. It allows

units to have control over the running variable, as the distribution of Z may depend on U in flexible ways. Yet, the condition $0 < F(0|u) < 1$ and the continuity of the conditional density ensure that such control may not be fully *precise* - i.e., it rules out *deterministic* sorting around the cutoff. For example, if for some u' we had $\Pr\{Z < 0|u'\} = 0$, then units with $U = u'$ would be all on one side of the cutoff and deterministic sorting would be possible - see Lee and Lemieux (2010a) for concrete examples.

Lee (2008, Proposition 2) shows that Assumption 2.2.1 implies the continuity condition in (2.2), is sufficient for identification of the ATE at the cutoff, and further implies that

$$H(w|z) \equiv \Pr\{W \leq w|Z = z\} \text{ is continuous in } z \text{ at } z = 0 \text{ for all } w \in \mathcal{W} \ . \tag{2.3}$$

In other words, the behavioral assumption that units do not precisely control Z around the cutoff implies that the treatment assignment is locally randomized *at the cutoff*, which means that the distribution of baseline covariates should not change discontinuously at the cutoff.

In this paper we propose a test for this null hypothesis of continuity in the distribution of the baseline covariates W at the cutoff $Z = 0$, i.e. (2.3). To better describe our test, it is convenient to define two auxiliary distributions that capture the local behavior of W to either side of the cutoff. To this end, define

$$H^-(w|0) = \lim_{z \uparrow 0} H(w|z) \quad \text{and} \quad H^+(w|0) = \lim_{z \downarrow 0} H(w|z) \ . \tag{2.4}$$

Using this notation, the continuity condition in (2.3) is equivalent to the requirement that $H(w|z)$ is right continuous at $z = 0$ and that

$$H^-(w|0) = H^+(w|0) \text{ for all } w \in \mathcal{W} \,. \tag{2.5}$$

The advantage of the representation in (2.5) is that it facilitates the comparison between two sample testing problems and the one we consider here. It also facilitates the comparison between our approach and alternative ones advocating the use of permutation tests on the grounds of favorable finite sample properties, see Section 2.3.2.

Remark 2.2.1. In randomized controlled experiments where the treatment assignment is exogenous by design, the empirical analysis usually begins with an assessment of the comparability of treated and control groups in baseline covariates, see Bruhn and McKenzie (2008). This practice partly responds to the concern that, if covariates differ across the two groups, the effect of the treatment may be confounded with the effect of the covariates - casting doubts on the validity of the experiment. The local randomization nature in RDD leads to the analogous (local) implication in (2.5). ■

Remark 2.2.2. Assumption 2.2.1 requires continuity of the conditional density of Z given U at $z = 0$, which implies continuity of the marginal density of Z, $f(z)$, at $z = 0$. McCrary (2008a) exploits this testable implication and proposes a test for the null hypothesis of continuity of $f(z)$ at the cutoff. Our test exploits a different implication of Assumption 2.2.1 and therefore should be viewed as a complement, rather than a substitute, to the density test proposed by McCrary (2008a). ■

Remark 2.2.3. Gerard et al. (2016) study the consequences of discontinuities in the density of Z at the cutoff. In particular, the authors consider a situation in which manipulation occurs only in one direction for a subset of the population (i.e., there exists a subset of participants such that $Z \geq 0$ a.s.) and use the magnitude of the discontinuity of $f(z)$ at $z = 0$ to identify the proportion of always-assigned units among all units close to the cutoff. Using this setup, Gerard et al. (2016) show that treatment effects in RDD are not point identified but that the model still implies informative bounds (i.e., treatment effects are partially identified). ■

A common practice in applied research is to test the hypothesis

$$E[W|Z = z] \text{ is continuous in } z \text{ at } z = 0 \ , \tag{2.6}$$

which is an implication of the null in (2.3). Table B.1 in Appendix B.0.5 shows that out of 62 papers published in leading journals during the period 2011-2015, 41 of them include a formal (or informal via some form of graphical inspection) test for the null in (2.6). However, if the fundamental hypothesis of interest is the implication derived by Lee (2008), testing the hypothesis in (2.6) has important limitations. First, tests designed for (2.6) have low power against certain distributions violating (2.3). Indeed, these tests may incorrectly lead the researcher to believe that baseline covariates are "continuous" at the cutoff, when some features of the distribution of W (other than the mean) may be discontinuous. Second, tests designed for (2.6) may exhibit poor size control in cases where usual smoothness conditions required for local polynomial estimation do not hold. Section 2.5 illustrates both of these points.

Before moving to describe the test we propose in this paper, we emphasize two aspects about Assumption 2.2.1 and the testable implication in (2.3). First, Assumption 2.2.1 is sufficient but not necessary for identification of the ATE at the cutoff. Second, the continuity condition in (2.3) is neither necessary nor sufficient for identification of the ATE at the cutoff. Assessing whether (2.3) holds or not is simply a sensible way to argue in favor or against the credibility of the design.

2.3. A permutation test based on induced ordered statistics

Let P be the distribution of (Y, W, Z) and $X^{(n)} = \{(Y_i, W_i, Z_i) : 1 \leq i \leq n\}$ be a random sample of n i.i.d. observations from P. Let q be a small (relative to n) integer. The test we propose is based on $2q$ values of $\{W_i : 1 \leq i \leq n\}$, such that q of these are associated with the q closest values of $\{Z_i : 1 \leq i \leq n\}$ to the right of the cutoff $\bar{z} = 0$, and the remaining q are associated with the q closest values of $\{Z_i : 1 \leq i \leq n\}$ to the left of the cutoff $\bar{z} = 0$. To be precise, denote by

$$Z_{n,(1)} \leq Z_{n,(2)} \leq \cdots \leq Z_{n,(n)} \tag{2.7}$$

the order statistics of the sample $\{Z_i : 1 \leq i \leq n\}$ and by

$$W_{n,[1]}, W_{n,[2]}, \ldots, W_{n,[n]} \tag{2.8}$$

the corresponding values of the sample $\{W_i : 1 \leq i \leq n\}$, i.e., $W_{n,[j]} = W_k$ if $Z_{n,(j)} = Z_k$ for $k = 1, \ldots, n$. The random variables in (2.8) are called *induced order statistics* or *concomitants* of order statistics, see David and Galambos (1974); Bhattacharya (1974).

In order to construct our test statistic, we first take the q closest values in (2.7) to the right of the cutoff and the q closest values in (2.7) to the left of the cutoff. We denote these ordered values by

$$Z^-_{n,(q)} \leq \cdots \leq Z^-_{n,(1)} < 0 \text{ and } 0 \leq Z^+_{n,(1)} \leq \cdots \leq Z^+_{n,(q)} \ , \tag{2.9}$$

respectively, and the corresponding induced values in (2.8) by

$$W^-_{n,[q]}, \ldots, W^-_{n,[1]} \text{ and } W^+_{n,[1]}, \ldots, W^+_{n,[q]} \ . \tag{2.10}$$

Note that while the values in (2.9) are ordered, those in (2.10) are not necessarily ordered.

The random variables $(W^-_{n,[1]}, \ldots, W^-_{n,[q]})$ are viewed as an independent sample of W conditional on Z being "close" to zero from the left, while the random variables $(W^+_{n,[1]}, \ldots, W^+_{n,[q]})$ are viewed as an independent sample of W conditional on Z being "close" to zero from the right. We therefore use each of these two samples to compute empirical cdfs as follows,

$$\hat{H}^-_n(w) = \frac{1}{q}\sum_{j=1}^{q} I\{W^-_{n,[j]} \leq w\} \text{ and } \hat{H}^+_n(w) = \frac{1}{q}\sum_{j=1}^{q} I\{W^+_{n,[j]} \leq w\} \ .$$

Finally, letting

$$S_n = (S_{n,1}, \ldots, S_{n,2q}) = (W^-_{n,[1]}, \ldots, W^-_{n,[q]}, W^+_{n,[1]}, \ldots, W^+_{n,[q]}) \ , \tag{2.11}$$

denote the pooled sample of induced order statistics, we can define our test statistic as

$$T(S_n) = \frac{1}{2q}\sum_{j=1}^{2q} (\hat{H}^-_n(S_{n,j}) - \hat{H}^+_n(S_{n,j}))^2 \ . \tag{2.12}$$

The statistic $T(S_n)$ in (2.12) is a Cramér Von Mises test statistic, see Hajek et al. (1999, p. 101).

We propose to compute the critical values of our test by a permutation test as follows. Let $\mathbf{G}$ denote the set of all permutations $\pi = (\pi(1), \ldots, \pi(2q))$ of $\{1, \ldots, 2q\}$. We refer to $\mathbf{G}$ as the group of permutations (in this context, "group" is understood as a mathematical group). Let

$$S_n^\pi = (S_{n,\pi(1)}, \ldots, S_{n,\pi(2q)}) \ ,$$

be the permuted values of S_n in (2.11) according to π. Let $M = |\mathbf{G}|$ be the cardinality of $\mathbf{G}$ and denote by

$$T^{(1)}(S_n) \leq T^{(2)}(S_n) \leq \cdots \leq T^{(M)}(S_n)$$

the ordered values of $\{T(S_n^\pi) : \pi \in \mathbf{G}\}$. For $\alpha \in (0,1)$, let $k = \lceil M(1-\alpha) \rceil$ and define

$$M^+(S_n) = |\{1 \leq j \leq M : T^{(j)}(S_n) > T^{(k)}(S_n)\}|$$

$$M^0(S_n) = |\{1 \leq j \leq M : T^{(j)}(S_n) = T^{(k)}(S_n)\}| \ , \tag{2.13}$$

where $\lceil x \rceil$ is the smallest integer greater than or equal to x. The test we propose is given by

$$\phi(S_n) = \begin{cases} 1 & T(S_n) > T^{(k)}(S_n) \\ a(S_n) & T(S_n) = T^{(k)}(S_n) \\ 0 & T(S_n) < T^{(k)}(S_n) \end{cases} \ , \tag{2.14}$$

where

$$a(S_n) = \frac{M\alpha - M^+(S_n)}{M^0(S_n)} \ .$$

Remark 2.3.1. The test in (2.14) is possibly randomized. The non-randomized version of the test that rejects when $T(S_n) > T^{(k)}(S_n)$ is also asymptotically level α by Theorem 2.4.2. In our simulations, the randomized and non-randomized versions perform similarly when M is not too small. ■

Remark 2.3.2. When M is too large the researcher may use a stochastic approximation to $\phi(S_n)$ without affecting the properties of our test. More formally, let

$$\hat{\mathbf{G}} = \{\pi_1, \ldots, \pi_B\} \ ,$$

where $\pi_1 = (1, \ldots, 2q)$ is the identity permutation and $\pi_2, \ldots, \pi_B$ are i.i.d. Uniform($\mathbf{G}$). Theorem 2.4.2 in Section 2.4 remains true if, in the construction of $\phi(S_n)$, $\mathbf{G}$ is replaced by $\hat{\mathbf{G}}$. ■

Remark 2.3.3. Our results are not restricted to the Cramér Von Mises test statistic in (2.12) and apply to other rank statistics satisfying our assumptions in Section 2.4, e.g., the Kolmogorov-Smirnov statistics. We restrict our discussion to the statistic in (2.12) for simplicity of exposition. ■

2.3.1. Implementing the new test

In this section we discuss the practical considerations involved in the implementation of our test, highlighting how we addressed these considerations in the companion `Stata` package `rdperm`.

The only tuning parameter of our test is the number q of observations closest to the cutoff. The asymptotic framework in Section 2.4 is one where q is fixed as $n \to \infty$, so this number should be small relative to the sample size. In this paper we recommend to use the following rule of thumb,

$$q_{\text{rot}} = \left\lceil f(0)\sigma_Z\sqrt{1-\rho^2}\frac{n^{0.9}}{\log n} \right\rceil , \tag{2.15}$$

where $f(0)$ is the density of Z at zero, ρ is the correlation between W and Z, and σ_Z^2 is the variance of Z. The motivation for this rule of thumb is as follows. First, the rate $\frac{n^{0.9}}{\log n}$ arises from the proof of Theorem 2.4.1, which suggests that q may increase with n as long as $n - q \to \infty$ and $q \log n/(n-q) \to 0$. Second, the constant arises by considering the special case where (W, Z) are bivariate normal. In such a case, it follows that

$$\left.\frac{\partial \Pr\{W \leq w | Z = z\}}{\partial z}\right|_{z=0} \propto \frac{-1}{\sigma_Z\sqrt{1-\rho^2}} \text{ at } w = E[W|z=0] .$$

Intuitively, one would like q to adapt to the slope of this conditional cdf. When the derivative is close to zero, a large q would be desired as in this case $H(w|0)$ and $H(w|z)$ should be similar for small values of $|z|$. When the derivative is high, a small value of q is desired as in this case $H(w|z)$ could be different than $H(w|0)$ even for small values of $|z|$. Our rule of thumb is thus inversely proportional to this derivative to capture this intuition. Finally, we scale the entire expression by the density of Z at the cutoff, $f(0)$, which accounts for the potential number of observations around the cutoff and makes q_{rot} scale invariant when (W, Z) are bivariate normal. All these quantities can be estimated to deliver a feasible $\hat{q}_{\text{rot}}$.

Given q, the implementation of our test proceeds in the following six steps.

Step 1. Compute the order statistics of $\{Z_i : 1 \leq i \leq n\}$ at either side of the cutoff as in (2.9).

Step 2. Compute the associated values of $\{W_i : 1 \leq i \leq n\}$ as in (2.10).

Step 3. Compute the test statistic in (2.12) using the observations from Step 2.

Step 4. Generate random permutations $\hat{\mathbf{G}} = \{\pi_1, \ldots, \pi_B\}$ as in Remark 2.3.2 for a given B.

Step 5. Evaluate the test statistic in (2.12) for each permuted sample: $T(S_n^{\pi_\ell})$ for $\ell \in \{1, \ldots, B\}$.

Step 6. Compute the p-value of the test as follows,

$$p_{\text{value}} = \frac{1}{B} \sum_{\ell=1}^{B} I\{T(S_n^{\pi_\ell}) \geq T(S_n)\} \ . \tag{2.16}$$

Note that p_{value} is the p-value associated with the non-randomized version of the test, see Remark 2.3.1. The default values in the Stata package, and the values we use in the simulations in Section 2.5, are $B = 999$ and $q = \hat{q}_{\text{rot}}$, as described in Appendix B.0.4.

Remark 2.3.4. The recommended choice of q in (2.15) is simply a sensible rule of thumb and is not an optimal rule in any formal sense. Given our asymptotic framework where q is fixed as n goes to infinity, it is difficult, and out of the scope of this paper, to derive optimal rules for choosing q. ■

Remark 2.3.5. The number of observations q on either side of the cutoff need not be symmetric. All our results go through with two fixed values, q_l and q_r, to the left and right of the cutoff respectively. However, we restrict our attention to the case where q

is the same on both sides as it simplifies deriving a rule of thumb for q and makes the overall exposition cleaner. ■

2.3.2. Relation to other permutation tests in the literature

Permutation tests have been previously discussed in the RDD literature for doing inference on distributional treatment effects. In particular, Cattaneo et al. (2015, CFT) provide conditions in a randomization inference context under which the RDD can be interpreted as a local randomized controlled experiment (RCE) and develop exact finite-sample inference procedures based on such an interpretation. Ganong and Jäger (2015) and Sales and Hansen (2015) build on the same framework and consider related tests for the kink design and projected outcomes, respectively.

The most important distinction with our paper is that permutation tests have been previously advocated on the grounds of finite sample validity. Such a justification requires, essentially, a different type of null hypothesis than the one we consider. In particular, suppose it was the case that for some $b > 0$, $H(w|z) = \Pr\{W \leq w | Z = z\}$ was constant in z for all $z \in [-b, b]$ and $w \in \mathcal{W}$. In other words, suppose the treatment assignment is locally randomized in a *neighborhood* of zero as opposed to "at zero". The null hypothesis in this case could be written as

$$H(w|z \in [-b, 0)) = H(w|z \in [0, b]) \text{ for all } w \in \mathcal{W} \,. \tag{2.17}$$

Under the null hypothesis in (2.17), a permutation test applied to the sample with observations $\{(W_i, Z_i) : -b \leq Z_i < 0\}$ and $\{(W_i, Z_i) : 0 \leq Z_i \leq b\}$, leads to a test that is valid in finite samples (i.e., its finite sample size does not exceed the nominal level). The

proof of this result follows from standard arguments (see Lehmann and Romano, 2005, Theorem 15.2.1). For these arguments to go through, the null hypothesis must be the one in (2.17) for a known b. Indeed, CFT clearly state that the key assumption for the validity of their approach is the existence of a *neighborhood* around the cutoff where a randomization-type condition holds. In our notation, this is captured by (2.17).

Contrary to those arguments, our paper shows that permutation tests can be used for the null hypothesis in (2.5), which only requires local randomization at zero, and shows that the justification for using permutation tests may be asymptotic in nature (see Remark 2.4.1 for a technical discussion). The asymptotics are non-standard as they intend to explicitly capture a situation where the number of effective observations (q in our notation) is small relative to the total sample size (n in our notation). This is possible in our context due to the recent asymptotic framework developed by Canay et al. (2017) for randomization tests, although we introduce novel modifications to make it work in the RDD setting - see Section 2.4.2. Therefore, even though the test we propose in this paper may be "mechanically" equivalent to the one in CFT, the formal arguments that justify their applicability are markedly different (see also the recent paper by Sekhon and Titiunik (2016) for a discussion on local randomization at the cutoff versus in a neighborhood). Importantly, while our test can be viewed as a test for (2.3), which is the actual implication in Lee (2008, Proposition 2), the test in CFT is a test for (2.17), which does not follow from Assumption 2.2.1.

Remark 2.3.6. The motivation behind the finite sample analysis in Cattaneo et al. (2015) is that only a few observations might be available close enough to the cutoff where a local randomization-type condition holds, and hence standard large-sample procedures

may not be appropriate. They go on to say that "...small sample sizes are a common phenomenon in the analysis of RD designs...", referring to the fact that the number of effective observations typically used for inference (those local to the cutoff) are typically small even if the total number of observations, n, is large. Therefore, the motivation behind their finite sample analysis is precisely the motivation behind our asymptotic framework where, as $n \to \infty$, the effective number of observations q that enter our test are taken to be finite. By embedding this finite sample situation into our asymptotic framework, we can construct tests for the hypothesis in (2.3) as opposed to the one in (2.17). ■

Remark 2.3.7. In Remark 2.2.1 we made a parallel between our testing problem and the standard practice in RCEs of comparing the treated and control groups in baseline covariates. However, the testable implication in RCEs is a global statement about the conditional distribution of W given $A = 1$ and $A = 0$. With large sample sizes, there exists a variety of asymptotically valid tests that are available to test $\Pr\{W \leq w|A = 1\} = \Pr\{W \leq w|A = 0\}$, and permutation tests are one of the many methods that may be used. On the contrary, in RDD the testable implication is "local" in nature, which means that few observations are actually useful for testing the hypothesis in (2.5). Finite sample issues, and permutation tests in particular, thus become relevant. ■

Another difference between the aforementioned papers and our paper is that their goal is to conduct inference on the (distributional) treatment effect and not on the hypothesis of continuity of covariates at the cutoff. Indeed, they essentially consider (sharp) hypotheses

of the form

$$Y_i(1) = Y_i(0) + \tau_i \text{ for all } i \text{ such that } Z_i \in [-b, b]$$

(for $\tau_i = 0 \; \forall i$ in the case of no-treatment effect), which deviates from the usual interest on average treatment effects (Ganong and Jäger, 2015, is about the kink design but similar considerations apply). On the contrary, the testable implication in Lee (2008, Proposition 2) is *precisely* a statement about conditional distribution functions (i.e. (2.3)), so our test is designed by construction for the hypothesis of interest.

Remark 2.3.8. Sales and Hansen (2015), building on CFT, also use small-sample justifications in favor of permutation tests. However, they additionally exploit the assumption that the researcher can correctly specify a model for variables of interest (outcomes in their paper and covariates in our setting) as a function of the running variable Z. Our results do not require such modeling assumptions and deliver a test for the hypothesis in (2.3) as opposed to (2.17). ■

Remark 2.3.9. Shen and Zhang (2016) also investigate distributional treatment effects in the RDD. In particular, they are interested in testing $\Pr\{Y(0) \leq y | Z = 0\} = \Pr\{Y(1) \leq y | Z = 0\}$, and propose a Kolmogorov-Smirnov-type test statistic based on local linear estimators of distributional treatment effects. Their asymptotic framework is standard and requires $nh \to \infty$ (where h is a bandwidth), which implies that the effective number of observations at the cutoff increases as the sample size increases. Although not mentioned in their paper, their test could be used to test the hypothesis in (2.3) whenever W is continuously distributed. We therefore compare the performance our test to the one in Shen and Zhang (2016) in Sections 2.5 and 2.6. ■

Remark 2.3.10. Our test can be used (replacing W with Y) to perform distributional inference on the outcome variable as in CFT and Shen and Zhang (2016). We do not focus on this case here. ■

2.4. Asymptotic framework and formal results

In this section we derive the asymptotic properties of the test in (2.14) using an asymptotic framework where q is fixed and $n \to \infty$. We proceed in two parts. We first derive a result on the asymptotic properties of induced order statistics in (2.10) that provides an important milestone in proving the asymptotic validity of our test. We then use this intermediate result to prove our main theorem.

2.4.1. A result on induced order statistics

Consider the order statistics in (2.7) and the induced order statistics in (2.8). As in the previous section, denote the q closest values in (2.7) to the right and left of the cutoff by

$$Z^-_{n,(q)} \leq \cdots \leq Z^-_{n,(1)} < 0 \text{ and } 0 \leq Z^+_{n,(1)} \leq \cdots \leq Z^+_{n,(q)} \ ,$$

respectively, and the corresponding induced values in (2.8) by

$$W^-_{n,[q]}, \ldots, W^-_{n,[1]} \text{ and } W^+_{n,[1]}, \ldots, W^+_{n,[q]} \ .$$

To prove the main result in this section we make the following assumption.

Assumption 2.4.1. *For any $\epsilon > 0$, Z satisfies* $\Pr\{Z \in [-\epsilon, 0)\} > 0$ *and* $\Pr\{Z \in [0, \epsilon]\} > 0$.

Assumption 2.4.1 requires that the distribution of Z is locally dense to the left of zero, and either locally dense to the right of zero or have a mass point at zero, i.e. $\Pr\{Z = 0\} > 0$. Importantly, Z could be continuous with a density $f(z)$ discontinuous at zero; or have mass points anywhere in the support except in a neighborhood to the left of zero.

Theorem 2.4.1. *Let Assumptions 2.4.1 and* (2.3) *hold. Then,*

$$\Pr\left\{\bigcap_{j=1}^{q}\{W_{n,[j]}^{-} \leq w_j^{-}\}\bigcap_{j=1}^{q}\{W_{n,[j]}^{+} \leq w_j^{+}\}\right\} = \Pi_{j=1}^{q} H^{-}(w_j^{-}|0) \cdot \Pi_{j=1}^{q} H^{+}(w_j^{+}|0) + o(1) \ ,$$

as $n \to \infty$*, for any* $(w_1^{-}, \ldots, w_q^{-}, w_1^{+}, \ldots, w_q^{+}) \in \mathbf{R}^{2q}$.

Theorem 2.4.1 states that the joint distribution of the induced order statistics are asymptotically independent, with the first q random variables each having limit distribution $H^{-}(w|0)$ and the remaining q random variables each having limit distribution $H^{+}(w|0)$. The proof relies on the fact the induced order statistics

$$S_n = (W_{n,[q]}^{-}, \ldots, W_{n,[1]}^{-}, W_{n,[1]}^{+}, \ldots, W_{n,[q]}^{+})$$

are conditionally independent given $(Z_1, \ldots, Z_n)$, with conditional cdfs

$$H(w|Z_{n,(q)}^{-}), \ldots, H(w|Z_{n,(1)}^{-}), H(w|Z_{n,(1)}^{+}), \ldots, H(w|Z_{n,(q)}^{+}) \ .$$

The result then follows by showing that $Z_{n,(j)}^{-} = o_p(1)$ and $Z_{n,(j)}^{+} = o_p(1)$ for all $j \in \{1, \ldots, q\}$, and invoking standard properties of weak convergence.

Theorem 2.4.1 plays a fundamental role in the proof of Theorem 2.4.2 in the next section. It is the intermediate step that guarantees that, under the null hypothesis in

(2.3), we have

$$S_n \xrightarrow{d} S = (S_1, \ldots, S_{2q}) \ , \tag{2.18}$$

where $(S_1, \ldots, S_{2q})$ are i.i.d. with cdf $H(w|0)$. This implies that $S^{\pi} \stackrel{d}{=} S$ for all permutations $\pi \in \mathbf{G}$, which means that the limit random variable S is indeed invariant to permutations.

Remark 2.4.1. Under the null hypothesis in (2.3) it is not necessarily true that the distribution of S_n is invariant to permutations. That is, $S_n^{\pi} \stackrel{d}{\neq} S_n$. Invariance of S_n to permutations is exactly the condition required for a permutation test to be valid in finite samples, see Lehmann and Romano (2005). The lack of invariance in finite samples lies behind the fact that the random variables in S_n are not draws from $H^-(w|0)$ and $H^+(w|0)$, but rather from $H(w|Z^-_{n,(j)})$ and $H(w|Z^+_{n,(j)})$, $j \in \{1, \ldots, q\}$. Under the null hypothesis in (2.3), the latter two distributions are not necessarily the same and therefore permuting the elements of S_n may not keep the joint distribution unaffected. However, under the continuity implied by the null hypothesis, it follows that a sample from $H(w|Z^-_{n,(j)})$ exhibits a similar behavior to a sample from $H^-(w|0)$, at least for n sufficiently large. This is the value of Theorem 2.4.1 to prove the results in the following section. ■

In addition to Assumptions 2.4.1, we also require that the random variable W is either continuous or discrete to prove the main result of the next section. Below we use $\text{supp}(\cdot)$ to denote the support of a random variable.

Assumption 2.4.2. *The scalar random variable W is continuously distributed conditional on $Z = 0$.*

Assumption 2.4.3. *The scalar random variable W is discretely distributed with $|\mathcal{W}| = m \in \mathbf{N}$ points of support and such that* $\text{supp}(W|Z = z) \subseteq \mathcal{W}$ *for all $z \in \mathcal{Z}$.*

We note that Theorem 2.4.1 does not require either Assumption 2.4.2 or Assumption 2.4.3. We however use each of these assumptions as a primitive condition of Assumptions 2.4.4 and 2.4.5 below, which are the high-level assumptions we use to prove the asymptotic validity of the permutation test in (2.14) for the scalar case. For ease of exposition, we present the extension to the case where W is a vector of possibly continuous and discrete random variables in Appendix B.0.3.

Remark 2.4.2. Our assumptions are considerably weaker than those used by Shen and Zhang (2016) to do inference on distributional treatment effects. In particular, while Assumption 2.4.1 allows Z to be discrete everywhere except in a local neighborhood to the left of zero, Shen and Zhang (2016, Assumption 3.1) require the density of Z to be bounded away from zero and twice continuously differentiable with bounded derivatives. Similar considerations apply to their conditions on $H(w|z)$. In addition, the test proposed by Shen and Zhang (2016) does not immediately apply to the case where W is discrete, as it requires an alternative implementation based on the bootstrap. On the contrary, our test applies indistinctly to continuous and discrete variables. ■

2.4.2. Asymptotic validity under approximate invariance

We now present our theory of permutation tests under approximate invariance. By approximate invariance we mean that only S is assumed to be invariant to $\pi \in \mathbf{G}$, while S_n may not be invariant - see Remark 2.4.1. The insight of approximating randomization tests when the conditions required for finite sample validity do not hold in finite samples, but are satisfied in the limit, was first developed by Canay et al. (2017) in a context where the group of transformations $\mathbf{G}$ was essentially sign-changes. Here we exploit this asymptotic framework but with two important modifications. First, our arguments illustrate a concrete case in which the framework in Canay et al. (2017) can be used for the group $\mathbf{G}$ of permutations as opposed to the group $\mathbf{G}$ of sign-changes. The result in Theorem 2.4.1 provides a fundamental milestone in this direction. Second, we adjust the arguments in Canay et al. (2017) to accommodate rank test statistics, which happen to be discontinuous and do not satisfy the so-called no-ties condition in Canay et al. (2017). We do this by exploiting the specific structure of rank test statistics, together with the requirement that the limit random variable S is either continuously or discretely distributed. We formalize our requirements for the continuous case in the following assumption, where we denote the set of distributions $P \in \mathbf{P}$ satisfying the null in (2.3) as

$$\mathbf{P}_0 = \{P \in \mathbf{P} : \text{condition (2.3) holds}\} .$$

Assumption 2.4.4. *If $P \in \mathbf{P}_0$, then*

(i) $S_n = S_n(X^{(n)}) \xrightarrow{d} S$ *under* P.

(ii) $S^\pi \overset{d}{=} S$ *for all* $\pi \in \mathbf{G}$.

(iii) S *is a continuous random variable taking values in* $\mathcal{S} \subseteq \mathbf{R}^{2q}$.

(iv) $T : \mathcal{S} \to \mathbf{R}$ *is invariant to rank, i.e., it only depends on the order of the elements in* $(S_1, \ldots, S_{2q})$.

Assumption 2.4.4 states the high-level conditions that we use to show the asymptotic validity of the permutation test we propose in (2.14) and formally state in Theorem 2.4.2 below. The assumption is also written in a way that facilitates the comparison with the conditions in Canay et al. (2017). In our setting, Assumption 2.4.4 follows from Assumptions 2.4.1-2.4.2, which may be easier to interpret and impose clear restrictions on the primitives of the model. To see this, note that Theorem 2.4.1, and the statement in (2.18) in particular, imply that Assumptions 2.4.4.(i)-(ii) follow from Assumption 2.4.1. In turn, Assumption 2.4.4.(iii) follows directly from Assumption 2.4.2. Finally, Assumption 2.4.4.(iv) holds for several rank test statistics and for the test statistic in (2.12) in particular.

To see the last point more clearly, it is convenient to write the test statistic in (2.12) using an alternative representation. Let

$$R_{n,i} = \sum_{j=1}^{2q} I\{S_{n,j} \leq S_{n,i}\} \ , \tag{2.19}$$

be the rank of $S_{n,i}$ in the pooled vector S_n in (2.11). Let $R^*_{n,1} < R^*_{n,2} < \cdots < R^*_{n,q}$ denote the increasingly ordered ranks $R_{n,1}, \ldots, R_{n,q}$ corresponding to the first sample (i.e., first q values) and $R^*_{n,q+1} < \cdots < R^*_{n,2q}$ denote the increasingly ordered ranks $R_{n,q+1}, \ldots, R_{n,2q}$ corresponding to the second sample (i.e., remaining q values). Letting

$$T^*(S_n) = \frac{1}{q}\sum_{i=1}^{q}(R^*_{n,i} - i)^2 + \frac{1}{q}\sum_{j=1}^{q}(R^*_{n,q+j} - j)^2 \tag{2.20}$$

it follows that

$$T(S_n) = \frac{1}{q}T^*(S_n) - \frac{4q^2-1}{12q} \ ,$$

see Hajek et al. (1999, p. 102). The expression in (2.20) immediately shows two properties of the statistic $T(s)$. First, $T(s)$ is not a continuous function of s as the ranks make discrete changes with s. Second, $T(s) = T(s')$ whenever s and s' share the same ranks (our Assumption 2.4.4(iv)), which immediately follows from the definition of $T^*(s)$. This property is what makes rank test statistics violate the no-ties condition in Canay et al. (2017).

We next formalize our requirements for the discrete case in the following assumption.

Assumption 2.4.5. *If $P \in \mathbf{P}_0$, then*

(i) $S_n = S_n(X^{(n)}) \xrightarrow{d} S$ *under* P.

(ii) $S^\pi \overset{d}{=} S$ *for all* $\pi \in \mathbf{G}$.

(iii) S_n *are discrete random variables taking values in* $\mathcal{S}_n \subseteq \mathcal{S} \equiv \otimes_{j=1}^{2q}\mathcal{S}_1$, *where* $\mathcal{S}_1 = \bigcup_{k=1}^{m}\{a_k\}$ *is a collection of* m *distinct singletons.*

Parts (i) and (ii) of Assumption 2.4.5 coincide with parts (i) and (ii) of Assumption 2.4.4 and, accordingly, follow from Assumption 2.4.1. Assumption 2.4.5.(iii) accommodates a case not allowed by Assumption 2.4.4.(iii), which required S to be continuous. This is important as many covariates are discrete in empirical applications, including the one in Section 2.6. Note that here we require the random variable S_n to be discrete, which in turn implies that S is discrete too. However, Assumption 2.4.5 does not impose any requirement on the test statistic $T : \mathcal{S} \to \mathbf{R}$.

We now formalize our main result in Theorem 2.4.2, which shows that the permutation test defined in (2.14) leads to a test that is asymptotically level α whenever either Assumption 2.4.4 or Assumption 2.4.5 hold. In addition, the same theorem also shows that Assumptions 2.4.1-2.4.3 are sufficient primitive conditions for the asymptotic validity of our test.

Theorem 2.4.2. *Suppose that either Assumption 2.4.4 or Assumption 2.4.5 holds and let $\alpha \in (0,1)$. Then, $\phi(S_n)$ defined in* (2.14) *satisfies*

$$E_P[\phi(S_n)] \to \alpha \tag{2.21}$$

as $n \to \infty$ whenever $P \in \mathbf{P}_0$. Moreover, if $T : \mathcal{S} \to \mathbf{R}$ is the test statistic in (2.12) *and Assumptions 2.4.1-2.4.2 hold, then Assumption 2.4.4 also holds and* (2.21) *follows. Additionally, if instead Assumptions 2.4.1 and 2.4.3 hold, then Assumption 2.4.5 also holds and* (2.21) *follows.*

Theorem 2.4.2 shows the validity of the test in (2.14) when the scalar random variable W is either discrete or continuous. However, the test statistic in (2.12) and the test construction in (2.14) immediately apply to the case where W is a vector consisting of a combination of discrete and continuously distributed random variables. In Appendix B.0.3 we show the validity of the test in (2.14) for the vector case, which is a result we use in the empirical application of Section 2.6. Also note that Theorem 2.4.2 implies that the proposed test is asymptotically similar, i.e., has limiting rejection probability equal to α if $P \in \mathbf{P}_0$.

Remark 2.4.3. If the distribution P is such that (2.17) holds and q is such that $-b \leq Z^-_{n,(q)} < Z^+_{n,(q)} \leq b$, then $\phi(S_n)$ defined in (2.14) satisfies

$$E_P[\phi_n(S_n)] = \alpha \text{ for all } n \ .$$

Since (2.17) implies (2.3), it follows that our test exhibits finite sample validity for some of the distributions in $\mathbf{P}_0$. ■

Remark 2.4.4. As in Canay et al. (2017), our asymptotic framework is such that the number of permutations in $\mathbf{G}$, $|\mathbf{G}| = q!$, is fixed as $n \to \infty$. An alternative asymptotic approximation would be one requiring that $|\mathbf{G}| \to \infty$ as $n \to \infty$ - see, for example, Hoeffding (1952), Romano (1989), Romano (1990), and more recently, Chung and Romano (2013) and Bugni, Canay and Shaikh (2016). This would require an asymptotically "large" number of observations local to the cutoff and would therefore be less attractive for the problem we consider here. From the technical point of view, these two approximations involve quite different formal arguments. ■

2.5. Monte Carlo Simulations

In this section, we examine the finite-sample performance of several different tests of (2.3), including the one introduced in Section 2.3, with a simulation study. The data for the study is simulated as follows. The scalar baseline covariate is given by

$$W_i = \begin{cases} m(Z_i) + U_{0,i} & \text{if } Z_i < 0 \\ m(Z_i) + U_{1,i} & \text{if } Z_i \geq 0 \end{cases}, \tag{2.22}$$

where the distribution of $(U_{0,i}, U_{1,i})$ and the functional form of $m(z)$ varies across specifications. In the baseline specification, we set $U_{0,i} = U_{1,i} = U_i$, where U_i is i.i.d. $N(0, 0.15^2)$, and use the same function $m(z)$ as in Shen and Zhang (2016), i.e.

$$m(z) = 0.61 - 0.02z + 0.06z^2 + 0.17z^3 \ .$$

The distribution of Z_i also varies across the following specifications.

Model 1: $Z_i \sim 2\text{Beta}(2,4) - 1$ where $\text{Beta}(a,b)$ is the Beta distribution with parameters (a,b).

Model 2: As in Model 1, but $Z_i \sim \frac{1}{2}(2\text{Beta}(2,8) - 1) + \frac{1}{2}(1 - 2\text{Beta}(2,8))$.

Model 3: As in Model 1, but values of Z_i with $Z_i \geq 0$ are scaled by $\frac{1}{4}$.

Model 4: As in Model 1, but Z_i is discretely distributed uniformly on the support

$$\{-1, -0.95, -0.90, \ldots, -0.15, -0.10, -\frac{3}{\sqrt{n}}, 0, 0.05, 0.10, 0.15, \ldots, 0.90, 0.95, 1\} \ .$$

Model 5: As in Model 1, but

$$m(z) = \begin{cases} 1.6 + z & \text{if } z < -0.1 \\ 1.5 - 0.4(z + 0.1) & \text{if } z \geq -0.1 \end{cases} \ .$$

Model 6: As in Model 5, but $Z_i \sim \frac{1}{2}(2\text{Beta}(2,8) - 1) + \frac{1}{2}(1 - 2\text{Beta}(2,8))$.

Model 7: as in Model 1, but

$$m(z) = \Phi\left(\frac{-0.85z}{1 - 0.85^2}\right) \ ,$$

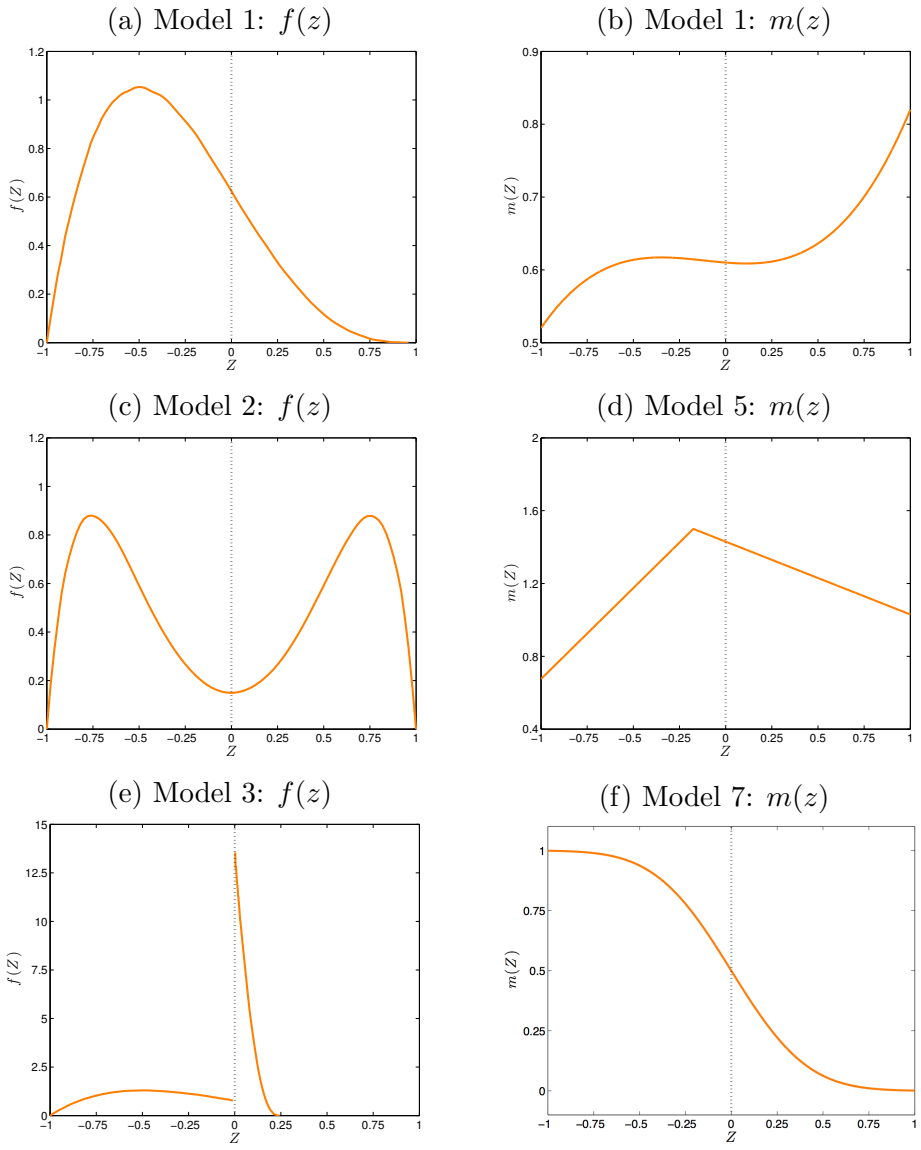


Figure 2.1. Density of Z (left column) and function $m(z)$ (right column) used in the Monte Carlo model specifications.

where $\Phi(\cdot)$ denotes the cdf of a standard normal random variable.

The baseline specification in Model 1 has two features: (i) Z_i is continuously distributed with a large number of observations around the cutoff; and (ii) the functional form of $m(z)$ is well behaved - differentiable and relatively flat around the cutoff, see Figure

2.1a and 2.1b. The other specifications deviate from the baseline as follows. Models 2 to 4 violate (i) in three different ways, see Figure 2.1c-2.1e. Model 5 violates (ii) by introducing a kink close to the cutoff, see Figure 2.1d. Model 6 combines Model 2 and 5 to violate both (i) and (ii). Finally, Model 7 is a difficult case (see Kamat, 2017, for a formal treatment of why this case is expected to introduce size distortions in finite samples) where the conditional mean of W exhibits a high first-order derivative at the threshold, see Figure 2.1f. These variations from the baseline model are partly motivated by the empirical application in Almond et al. (2010), where the running variable may be viewed as discrete as in Model 4, having heaps as in Figure 2.1c, or exhibiting discontinuities as in Figure 2.1e.

We consider sample sizes $n \in \{1000, 2500, 5000\}$, a nominal level of $\alpha = 5\%$, and perform $10,000$ Monte Carlo repetitions. Models 1 to 7 satisfy the null hypothesis in (2.3). We additionally consider the same models but with $U_{0,i} \overset{d}{\neq} U_{1,i}$ to examine power under the alternative.

Model P1-P7: Same as Models 1-7, but $U_{1,i} \sim \frac{1}{2}N\left(0.2, 0.15^2\right) + \frac{1}{2}N\left(-0.2, 0.15^2\right)$.

We report results for the following tests.

RaPer and **Per**: the permutation test we propose in this paper in its two versions. The randomized version (RaPer) in (2.14) and the non-randomized version (Per) that rejects when p_{value} in (2.16) is below α, see Remark 2.3.1. We include the randomized version only in the results on size to illustrate the differences between the randomized and non-randomized versions of the test. For power results, we simply report Per, which is the version of the test that practitioners

will most likely use. The tuning parameter q is set to

$$q \in \{10, 25, 50, q_{\text{rot}}, \hat{q}_{\text{rot}}\} \ ,$$

where q_{rot} is the rule of thumb in (2.3.4) and $\hat{q}_{\text{rot}}$ is a feasible q_{rot} with all unknown quantities non-parametrically estimated - see Appendix B.0.4 for details. We set $B = 999$ for the random number of permutations, see Remark 2.3.2.

SZ: the test proposed by Shen and Zhang (2016) for the null hypothesis of no distributional treatment effect at the cutoff. When used for the null in (2.3) at $\alpha = 5\%$, this test rejects when

$$A\left(\frac{n}{2}\tilde{f}_n\right)^2 \sup_w \left|\tilde{H}_n^-(w) - \tilde{H}_n^+(w)\right| \ ,$$

exceeds 1.3581. Here A is a known constant based on the implemented kernel, $\tilde{f}_n$ is a nonparametric estimate of the density of Z_i at $Z_i = 0$, and $\tilde{H}_n^-(w)$ and $\tilde{H}_n^+(w)$ are local linear estimates of the cdfs in (2.4). The kernel is set to a triangular kernel. Shen and Zhang (2016) propose using the following (undersmoothed) rule of thumb bandwidth for the nonparametric estimates,

$$h_n = h_n^{CCT} n^{1/5 - 1/c_h} \ , \tag{2.23}$$

where h_n^{CCT} is a sequential bandwidth based on Calonico et al. (2014b), and c_h is an undersmoothing parameter - see Appendix B.0.4 for details. We follow Shen and Zhang (2016) and report results for $c_h \in \{4.0, 4.5, 5.0\}$, where $c_h = 4.5$ is their recommended choice.

Model	n	RaPer					Per					SZ			CCT
		q					q					c_h			
		10	25	50	q_{rot}	$\hat{q}_{\text{rot}}$	10	25	50	q_{rot}	$\hat{q}_{\text{rot}}$	4.0	4.5	5.0	
	1000	5.18	4.83	4.92	4.85	4.89	5.05	4.82	4.92	4.79	4.87	3.86	4.29	5.03	5.54
1	2500	4.67	5.08	4.86	4.81	4.76	4.57	5.06	4.85	4.80	4.75	4.10	4.70	5.44	4.90
	5000	5.34	5.23	4.75	4.57	4.53	5.24	5.21	4.75	4.56	4.53	4.02	4.54	5.08	4.28
	1000	5.17	5.31	4.98	5.06	5.10	5.04	5.30	4.97	4.94	4.99	3.89	5.49	7.34	6.36
2	2500	5.15	5.02	5.01	5.09	4.85	5.01	5.02	5.00	5.03	4.77	4.37	6.13	8.65	5.39
	5000	5.02	5.35	4.92	5.18	5.35	4.93	5.34	4.92	5.16	5.34	4.55	6.23	9.03	4.87
	1000	5.17	4.86	4.90	4.86	4.77	5.05	4.84	4.90	4.80	4.77	13.54	13.84	13.97	7.80
3	2500	4.67	5.06	4.82	4.78	4.74	4.58	5.05	4.82	4.77	4.74	12.60	12.85	13.31	5.90
	5000	5.35	5.23	4.74	4.60	4.64	5.25	5.21	4.73	4.59	4.64	13.53	13.73	14.06	5.40
	1000	4.84	4.63	4.69	4.93	5.02	4.75	4.62	4.69	4.82	5.01	26.80	18.21	15.00	3.94
4	2500	5.09	5.06	5.00	4.92	4.96	5.00	5.05	5.00	4.91	4.96	16.19	11.73	10.52	4.50
	5000	4.59	5.01	4.76	4.98	4.80	4.53	4.98	4.76	4.97	4.80	7.66	7.18	7.75	5.30
	1000	5.37	6.18	17.29	5.43	5.49	5.27	6.16	17.27	5.32	5.38	4.93	8.23	13.00	5.71
5	2500	4.66	5.34	6.71	5.02	5.11	4.54	5.34	6.71	4.99	5.08	6.73	14.36	25.22	5.05
	5000	5.38	5.34	5.32	5.19	5.05	5.30	5.32	5.32	5.19	5.05	8.64	21.29	35.04	3.97
	1000	6.77	18.20	16.50	6.81	6.85	6.62	18.15	16.50	6.61	6.74	7.30	13.91	21.02	10.19
6	2500	5.67	10.00	33.07	5.65	5.62	5.58	9.98	33.05	5.53	5.60	12.02	26.10	39.26	12.17
	5000	5.03	6.48	14.40	5.91	6.43	4.89	6.46	14.38	5.91	6.42	16.94	40.55	56.54	13.55
	1000	6.03	19.74	85.07	5.99	5.98	5.94	19.70	85.07	5.88	5.86	5.10	7.05	9.86	5.69
7	2500	4.83	7.22	24.08	5.88	5.68	4.71	7.22	24.05	5.84	5.64	5.02	6.97	10.14	5.06
	5000	5.33	6.08	9.25	6.35	6.34	5.26	6.05	9.24	6.35	6.34	4.83	6.36	9.08	4.26

Table 2.1. Rejection probabilities (in %) under the null hypothesis. 10,000 replications.

CCT: the test proposed by Calonico et al. (2014b) for the null hypothesis of no average treatment effect at the cutoff. When used for the null in (2.3) at $\alpha = 5\%$, this test rejects when

$$\frac{\left|\hat{\mu}_n^{-,bc} - \hat{\mu}_n^{+,bc}\right|}{\hat{V}_n^{bc}},$$

exceeds 1.96. Here $\hat{\mu}_n^{-,bc}$ and $\hat{\mu}_n^{+,bc}$ are bias corrected local linear estimates of the conditional means of W_i to the left and right of $Z_i = 0$, and $\hat{V}_n^{bc}$ is a novel standard error formula that accounts for the variance of the estimated bias. The kernel is set to a triangular kernel. We implement their test using their proposed bandwidth - see Appendix B.0.4 for details.

Model	n	Per					SZ			SZ (Size Adj.)			CCT
		q					c_h			c_h			
		10	25	50	q_{rot}	$\hat{q}_{\text{rot}}$	4.0	4.5	5.0	4.0	4.5	5.0	
P1	1000	8.23	19.20	52.62	12.77	12.04	13.80	18.71	23.75	17.36	20.98	23.59	5.93
	2500	8.46	21.17	53.76	30.39	30.15	40.95	55.67	67.06	45.58	57.09	65.13	4.72
	5000	8.43	20.07	53.05	60.70	60.53	81.70	92.50	96.90	86.19	93.68	96.77	4.97
P2	1000	8.73	20.17	53.10	8.80	8.69	7.17	11.28	17.19	8.90	10.33	11.88	7.40
	2500	8.38	19.22	52.69	10.41	11.24	17.18	29.61	44.97	19.18	26.11	30.92	5.47
	5000	8.24	20.45	53.74	18.57	21.00	42.75	65.16	81.28	44.86	59.10	69.75	5.19
P3	1000	8.23	19.20	52.59	12.56	20.89	16.08	19.38	22.23	4.48	5.52	6.58	6.47
	2500	8.44	21.17	53.84	30.25	59.68	34.92	43.00	51.26	13.30	17.40	22.73	5.29
	5000	8.43	20.05	52.96	60.81	92.56	66.50	78.38	86.36	33.33	46.69	57.67	4.95
P4	1000	8.16	20.58	53.92	8.70	15.85	43.75	38.91	41.33	5.33	7.95	12.04	4.59
	2500	8.41	20.08	52.85	16.12	41.59	61.45	71.04	80.12	15.88	36.91	57.09	4.87
	5000	8.52	20.50	53.36	33.62	78.25	91.78	97.52	99.05	84.72	94.94	97.96	4.83
P5	1000	8.40	20.43	56.84	9.47	9.43	15.42	22.01	29.66	15.54	15.05	14.27	5.84
	2500	8.46	21.18	53.86	19.83	20.39	39.59	53.19	65.70	32.20	26.11	20.93	4.79
	5000	8.55	20.25	52.99	40.69	41.58	70.74	83.28	90.95	55.88	37.09	23.39	4.81
P6	1000	9.24	25.94	46.50	9.24	9.16	11.36	20.53	31.43	8.14	9.02	10.62	9.37
	2500	8.68	21.89	62.01	10.42	10.89	20.26	34.00	46.84	10.40	10.70	12.46	10.02
	5000	8.16	21.57	56.69	17.85	19.14	30.25	44.32	57.73	13.02	12.03	13.65	12.72
P7	1000	8.89	26.94	81.03	8.81	9.01	16.58	24.87	32.98	16.27	18.04	18.17	5.92
	2500	8.48	21.77	58.89	16.06	16.02	48.57	66.94	80.30	48.40	57.06	58.06	4.78
	5000	8.53	20.33	54.33	31.83	31.11	85.00	95.62	98.64	85.56	93.38	95.72	4.85

Table 2.2. Rejection probabilities (in %) under the alternative hypothesis. 10,000 replications.

Table 2.1 reports rejection probabilities under the null hypothesis for all models and all tests considered. Across all cases, the permutation test controls size remarkably well. In particular, the feasible rule of thumb $\hat{q}_{\text{rot}}$ in (2.3.4) delivers rejection rates between 4.53% and 6.74%. On the other hand, SZ returns rejection rates between 4.29% and 40.55% for their recommended choice of $c_h = 4.5$. Except in the baseline Model 1 where SZ performs similarly to Per, in all other models Per clearly dominates SZ in terms of size control. Finally, CCT controls size very well in all models except Model 6, where the lack of smoothness affects the local polynomial estimators and returns rejection rates between

Model	n	Per		SZ			CCT
		q		c_h			
		q_{rot}	$\hat{q}_{\text{rot}}$	4.0	4.5	5.0	
	1000	17.00	16.59	90.76	109.95	128.11	137.48
1	2500	33.00	32.93	219.93	273.20	324.91	349.21
	5000	56.00	56.08	427.10	540.87	653.45	699.11
	1000	10.00	10.00	49.60	67.43	88.09	98.84
2	2500	14.00	14.93	120.33	170.24	230.87	255.08
	5000	23.00	24.52	230.78	335.18	466.67	500.58
	1000	17.00	25.91	62.50	73.20	82.61	86.50
3	2500	33.00	54.23	147.40	176.64	202.83	213.31
	5000	56.00	95.59	283.75	346.32	402.97	423.25
	1000	11.00	19.91	116.78	141.49	165.08	179.68
4	2500	21.00	40.58	260.21	324.02	385.11	415.71
	5000	36.00	69.88	494.24	624.02	755.82	801.74
	1000	12.00	11.89	94.85	114.90	133.89	124.83
5	2500	23.00	23.48	222.03	275.94	328.25	281.12
	5000	40.00	39.89	401.72	508.99	614.76	487.69
	1000	10.00	10.00	51.20	69.97	91.80	89.92
6	2500	12.00	13.60	115.61	163.21	221.12	199.77
	5000	21.00	22.30	208.29	300.11	414.91	324.12
	1000	10.00	10.05	94.81	114.86	133.90	132.01
7	2500	18.00	18.42	203.43	252.76	300.72	319.24
	5000	31.00	31.22	347.02	439.56	531.06	604.18

Table 2.3. Average number of observations (to one side) used in the tests reported in Table 2.1.

10.19% and 13.55%. Table 2.3 reports the average number of observations[4] used by each of the tests and illustrates how both SZ and CCT consistently use a larger number of observations around the cutoff than Per.

Two final lessons arise from Table 2.1. First, the differences between RaPer and Per are negligible, even when $q = 10$. Second, Per is usually less sensitive to the choice of q than SZ is to the choice of c_h. The notable exceptions are Model 6, where both tests appear to be equally sensitive; and Model 7, where Per is more sensitive for $n = 1,000$

[4]In the case of SZ and CCT, we compute the average of the number of observations to the left and right of the cutoff, and then take an average across simulations. In the case of Per, we simply average q across simulations.

and $n = 2,500$. Recall that Model 7 is a particularly difficult case in RDD (see Kamat, 2017), but even in this case Per controls size well for n sufficiently large or q sufficiently small. Most importantly, the rejection probabilities under the null hypothesis are very close to the nominal level for our suggested rule of thumb $\hat{q}_{\text{rot}}$.

Table 2.2 reports rejection probabilities under the alternative hypothesis for all models and all tests considered. Since SZ may severely over-reject under the null hypothesis, we report both raw and size-adjusted rejection rates. For the recommended values of tuning parameters, the size adjusted power of SZ is consistently above the one of Per in Models P1, P2, and P7. In Models P3-P6, Per delivers higher power than SZ in 9 out of the 12 cases considered; while in the remaining three cases (P4 with $n = 5,000$ and P5 with $n \in \{2,500, 5,000\}$), SZ delivers higher power. This is remarkable as Table 2.3 shows that Per uses considerably fewer observations than SZ does.[5] The power of CCT, as expected, does not exceed the rejection probabilities under the null hypothesis.

2.6. Empirical application

In this section we reevaluate the validity of the design in Lee (2008). Lee studies the benefits of incumbency on electoral outcomes using a discontinuity constructed with the insight that the party with the majority wins. Specifically, the running variable Z is the difference in vote shares between Democrats and Republicans in time t. The assignment rule then takes a cutoff value of zero that determines the treatment of incumbency to the Democratic candidate, which is used to study their election outcomes in time $t + 1$. The data set contains six covariates that contain electoral information

[5]We computed the equivalent of Table 2.3 for the results in Table 2.2 and obtained very similar numbers, so we only report Table 2.3 to save space.

on the Democrat runner and the opposition in time $t-1$ and t. Out of the six variables, one is continuous (Democrat vote share $t-1$) and the remaining are discrete. The total number of observations is 6,559 with 2,740 below the cutoff. The dataset is publicly available at `http://economics.mit.edu/faculty/angrist/data1/mhe` and all the results in this section were computed using the `rdperm Stata` package available at `http://sites.northwestern.edu/iac879/software/`.

Lee assessed the credibility of the design in this application by inspecting discontinuities in means of the baseline covariates. His test is based on local linear regressions with observations in different margins around the cutoff. The estimates and graphical illustrations of the conditional means are used to conclude that there are no discontinuities at the cutoff in the baseline covariates. Here, we frame the validity of the design in terms of the hypothesis in (2.3) and use the newly developed permutation test as described in Section 2.3.1, using $\hat{q}_{\text{rot}}$ as our default choice for the number of observations q.[6] Our test allows for continuous or discrete covariates, and so it does not require special adjustments to accommodate discrete covariates; cf. Remark 2.4.2. In addition, our test allows the researcher to test for the hypothesis of continuity of individual covariates, in which case W includes a single covariate; as well as continuity of the entire vector of covariates, in which case W includes all six covariates. Finally, we also report the results of test CCT, as described in Section 2.5, for the continuity of means at the cutoff.

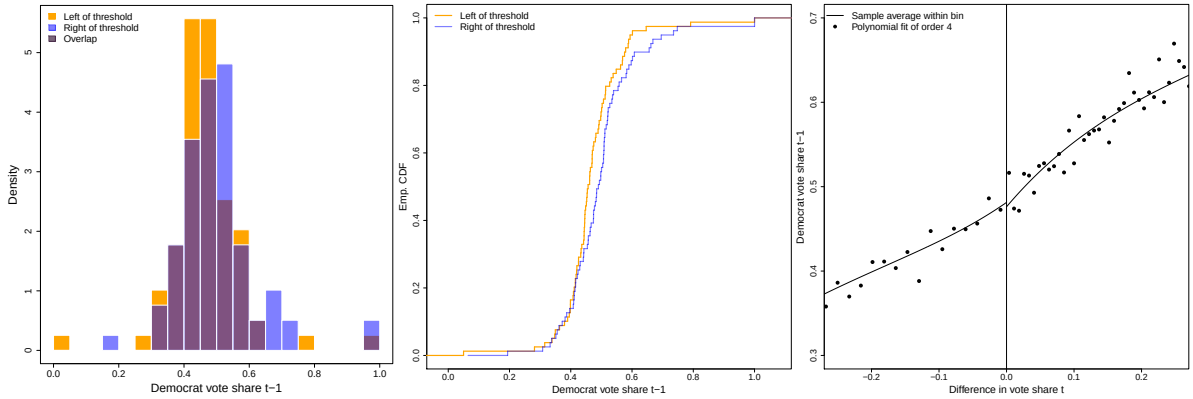
Table 2.4 reports the p-values for continuity of each of the six covariates individually, as well as the joint test for the continuity of the six dimensional vector of covariates; see Appendix B.0.3 for details. Our results show that the null hypothesis of continuity of the

[6]We also computed our test using $0.75\hat{q}_{\text{rot}}$ and $1.25\hat{q}_{\text{rot}}$ and found similar results.

Variable	Per	CCT	SZ
Democrat vote share $t-1$	4.60	83.74	31.21
Democrat win $t-1$	1.20	7.74	–
Democrat political experience t	0.30	21.43	–
Opposition political experience t	3.60	83.14	–
Democrat electoral experience t	13.31	25.50	–
Opposition electoral experience t	4.20	92.79	–
Joint Test - CvM statistic	16.42		
Joint Test - Max statistic	1.70		

Table 2.4. Test results with p-value (in %) for covariates in Lee (2008)

conditional distributions of the covariates at the cutoff is rejected for most of the covariates at a 5% significance level, in contrast to the results reported by Lee (2008) and the results of the CCT test in Table 2.4. The differences between our test and tests based on conditional means can be illustrated graphically. Figure 2.2(a)-(b) displays the histogram and empirical cdf (based on $\hat{q}_{\text{rot}}$ observations on each side) of the continuous covariate *Democrat vote share* $t-1$. The histogram exhibits a longer right tail for observations to the right of the threshold (in orange), and significantly more mass at shares below 50% for observations to the left of the threshold (in blue). The empirical CDFs are similar up until the 20th quantile, approximately, and then are markedly different. Our test formally shows that the observed differences are statistically significant. On the contrary, the conditional means from the left and from the right appear to be similar around the cutoff and so tests for the null hypothesis in (2.6) fail to reject the null in (2.3); see Figure 2.2c. A similar intuition applies to the rest of the covariates. Finally, we note that $\hat{q}_{\text{rot}}$ in the implementation of our test ranges from 51 to 310, depending on the covariate, while the average number of effective observations (i.e. the average of observations to the left and right of the cutoff) used by CCT ranges from 828 to 1192. This is consistent with one



(a) Histogram (b) CDF (c) Conditional Mean

Figure 2.2. Histogram, CDF, and conditional means for *Democrat vote share* $t-1$

asymptotic framework assuming few effective observations around the cutoff and another assuming a large and growing number of observations around the cutoff.

Table 2.4 also reports the test by Shen and Zhang (2016), as described in Section 2.5, for the only continuously distributed covariate of this application. This tests fails to reject the null hypothesis with a p-value of 31.21%. The rest of the covariates in this empirical application are discrete and so the results in Shen and Zhang (2016) do not immediately apply; see Remark 2.4.2 and Appendix B.0.4 for details.

The standard practice in applied work appears to be to test the hypothesis of continuity individually for each covariate. This is informative as it can provide information as to which covariate may or may not be problematic. However, testing many individual hypotheses may lead to spurious rejections (due to a multiple testing problem). In addition, the statement in (2.3) is a statement about the vector W that includes all baseline

covariates in the design. We therefore report in Table 2.4, in addition to each individual test, the results for the joint test that uses all six covariates in the construction of the test statistic - as explained in detail in Section B.0.3. Table 2.4 shows that the results for the joint test depend on the choice of test statistic used in its construction. If one uses the Cramér Von Mises test statistic in (2.12), the null hypothesis in (2.3) is not rejected, with a p-value of 17.62%. If one instead uses the max-type test statistic introduced in Appendix B.0.3, see (B.11), the null hypothesis in (2.3) is rejected, with a p-value of 1.70%. In unreported simulations we found that the max-type test statistic appears to have significantly higher power than the Cramér Von Mises test statistic in the multivariate case, which is consistent with the results of this particular application. It is worth noting that in the case of scalar covariates, these two test statistics are numerically identical. We therefore recommend the Max test statistic in (B.11) for the multivariate case, which is the default option in the companion `rdperm Stata` package.

2.7. Concluding remarks

In this paper we propose an asymptotically valid permutation test for the hypothesis of continuity of the distribution of baseline covariates at the cutoff in the regression discontinuity design (RDD). The asymptotic framework for our test is based on the simple intuition that observations close to the cutoff are *approximately* identically distributed on either side of it when the null hypothesis holds. This allows us to permute these observations to conduct an approximately valid test. Formally, we exploit the framework, with novel additions, from Canay et al. (2017), which first developed the insight of approximating randomization tests in this manner. Our results also represent a novel application

of induced order statistics to frame our problem, and we present a result on induced order statistics that may be of independent interest.

A final aspect we would like to highlight of our test is its simplicity. The test only requires computing two empirical cdfs for the induced order statistic, and does not involve kernels, local polynomials, bias correction, or bandwidth choices. Importantly, we have developed the rdperm Stata package that allows for effortless implementation of the test we propose in this paper.

CHAPTER 3

On Nonparametric Inference in the Regression Discontinuity Design

3.1. Introduction

The nonparametric literature on the regression discontinuity design (RDD) is characterized by the nonparametric identification of parameters at the threshold. In this paper we study constructing tests for these parameters, for which numerous alternatives are present in econometrics - see, for example, McCrary (2008b), Frandsen et al. (2012), Calonico et al. (2014a) and Otsu et al. (2015) for such tests, and see Lee and Lemieux (2010b) and Imbens and Lemieux (2008b) for recent surveys on the literature. In particular, we focus on the null hypothesis that the average treatment effect at the threshold in the sharp design equals a pre-specified value.

When testing this null hypothesis in simulation studies (not reported), we observe that available tests fail to control the rejection probability under some null distributions with practical sample sizes. This failure occurs for distributions that satisfy the typically imposed assumptions, and in turn makes us question the reliability of current inference procedures. Here we hence formally study the construction of valid tests for our null hypothesis. As stated in Section 3.3, the aim is to ideally construct a finite sample valid test, which requires the finite sample control of size, i.e. the null rejection probabilities. Since nontrivially achieving this may be too demanding, one may aim to approximate this

finite sample goal in large samples through two different definitions of asymptotic validity. The first termed uniform asymptotic validity requires limiting control of null rejection probability uniformly across distributions under the null, whereas the second termed pointwise asymptotic validity requires such control to hold for each fixed distribution under the null. As highlighted in Remark 3.4.3, current tests are shown to only satisfy the second definition, which may not provide any guarantee on the control of finite sample size. The practical importance of the distinction in these definitions has also been previously noted in various other econometric applications - see, for example, Mikusheva (2007) and Mikusheva (2012) for unit roots in autoregressive models, Romano and Shaikh (2008b) and Andrews and Guggenberger (2009b) for moment inequality models, Leeb and Pötscher (2005) and Andrews and Guggenberger (2009a) for post model selection and Dufour (1997) and Mikusheva (2010) for weak instrumental variable models.

Our first result establishes that, under standard assumptions in the basic setup, any test for our null hypothesis of interest will have power against any alternative bounded above by its size. This implies that it is impossible to construct nontrivial finite sample valid tests and uniformly asymptotically valid tests under these assumptions. Intuitively, this result occurs because the assumptions permit a set of possible distributions that is 'too large', in a sense made precise in Lemma 3.3.1. This causes distributions under the null and alternative to be 'arbitrarily' close making it impossible to distinguish them given the data. Our goal through this impossibility result is not to criticize current nonparametric tests but to attempt to caution researchers using them. Such nonparametric tests are often viewed as appealing as they only require imposing mild regularity assumptions. We hope to convey that these assumptions however allow the permitted set of distributions

to be arbitrarily large resulting in misleading inference. To recover reliable inference, the researcher would then naturally need to strengthen the assumptions to further restrict the permitted set of distributions. To this end, our second result illustrates a sufficient strengthening of the standard assumptions under which a version of the Calonico et al. (2014a) test is uniformly asymptotically valid. Our stronger assumptions are analogous to the ones commonly required for optimality results in nonparametric estimation; see, for example, van der Vaart (1998b, Chapter 24).

In addition to the literature on RDD, this paper is also related to the growing one in econometrics on the testability of hypotheses. Bahadur and Savage (1956) was the initial paper to demonstrate the impossibility of constructing nontrivial valid tests for the mean of a distribution. Romano (2004) extended this result to provide sufficient conditions to examine the testability of hypotheses in different settings. The key insight is formalizing the notion of closeness of the set of null and alternative distributions using the total variation metric. In econometrics, Canay et al. (2013) verified one of these conditions to establish impossibility of constructing nontrivial valid tests for some hypotheses in nonparametric models with endogeneity. In this paper we verify the same condition, restated as Lemma 3.3.1 here, to prove our impossibility result. Alternatively, Guggenberger (2010a,b) used a direct approach of considering sequences of distributions under the null to show limiting size distortions in the Hausman pretest. For further examples of such impossibility results see Lehmann and Loh (1990), Leeb and Pötscher (2008) and Müller (2008), and for a review of such results in econometrics see Dufour (2003).

The remainder of the paper is organized as follows. Section 3.2 describes the basic RDD setup, where we introduce the notation, the commonly imposed assumptions and the

null hypothesis of interest. Section 3.3 states our testing problem. Section 3.4 illustrates our main results.

3.2. Basic RDD Setup

Assume there are random variables $(Y(0), Y(1), Z) \sim Q \in \mathbf{Q}$, where $\mathbf{Q}$ is a set of distributions on a sample space $\mathcal{W} = \mathcal{Y} \times \mathcal{Y} \times \mathcal{Z} \subseteq \mathbf{R} \times \mathbf{R} \times \mathbf{R}$ such that $\mathcal{Z}$ contains a neighbourhood of zero. Here, let $Y(0)$ denote the potential outcome under treatment zero, $Y(1)$ denote the potential outcome under treatment one, and Z denote an observed predetermined characteristic. The observed random variables from the experiment are $(Y, Z) \sim P \in \mathbf{P}$, where $\mathbf{P}$ is a set of distributions on a sample space $\mathcal{X} = \mathcal{Y} \times \mathcal{Z} \subseteq \mathbf{R} \times \mathbf{R}$. The observed outcome is determined by

$$Y = A \cdot Y(1) + (1 - A) \cdot Y(0) \ , \tag{3.1}$$

where treatment assignment follows a normalized threshold rule of the form

$$A = 1\{Z \geq 0\} \ . \tag{3.2}$$

Since

$$(Y, Z) = M(Y(0), Y(1), Z) \ , \tag{3.3}$$

where $M : \mathcal{W} \to \mathcal{X}$ is the mapping implied by (3.1), we have that $P = QM^{-1}$ and

$$\mathbf{P} = \{QM^{-1} : Q \in \mathbf{Q}\} \ , \tag{3.4}$$

where M^{-1} is the pre-image of M. Let $W^{(n)} = \{(Y_i(0), Y_i(1), Z_i) : 1 \leq i \leq n\}$ denote an i.i.d sample from Q, and let $X^{(n)} = \{(Y_i, Z_i) : 1 \leq i \leq n\}$ denote the corresponding observed i.i.d sample from P. Further, let P^n denote the n-fold product $\bigotimes_{i=1}^{n} P$, i.e. the joint distribution of the observed data.

We next illustrate the standard assumptions and the resulting set of possible distributions $\mathbf{Q}$, which plays a fundamental role in our analysis. In order to do so, we introduce further notation. Let $\mu_-(z, Q) = E_Q[Y(0)|Z = z]$ and $\mu_+(z, Q) = E_Q[Y(1)|Z = z]$, and, whenever $\mu_-(\cdot, Q)$ and $\mu_+(\cdot, Q)$ have the appropriate level of differentiability, let $\mu_-^v(z, Q) = d^v\mu_-(z, Q)/dz^v$ and $\mu_+^v(z, Q) = d^v\mu_+(z, Q)/dz^v$. Further, let $\sigma_-^2(z, Q) = Var_Q[Y(0)|Z = z]$ and $\sigma_+^2(z, Q) = Var_Q[Y(1)|Z = z]$, and let $f_Q(z)$ denote the density of Z. Using this notation, let

$$\mathbf{Q} = \{Q \in \mathbf{Q}_\mathcal{W} : Q \text{ satisfies Assumption 3.2.1}\} , \tag{3.5}$$

where $\mathbf{Q}_\mathcal{W}$ denotes the set of all Borel probability measures on $\mathcal{W}$ that have a density on Z with respect to the Lebesgue measure, and Assumption 3.2.1 is stated below. Assumption 3.2.1, in particular, captures the commonly imposed restrictions in the majority of the nonparametric RDD literature; see, for example, Calonico et al. (2014a) and Imbens and Kalyanaraman (2012b).

Assumption 3.2.1. *Let Q be such that there exist real numbers $\kappa(Q) > 0$, $L(Q) > 0$ and $U(Q) > 0$ where for all $z \in (-\kappa(Q), \kappa(Q))$, i.e. in a neighbourhood around the threshold, the following conditions hold true.*

(i) $f_Q(z)$ is continuous and $L(Q) \leq f_Q(z) \leq U(Q)$.

(ii) $E_Q\left[Y(0)^4|Z=z\right] \leq U(Q)$ *and* $E_Q\left[Y(1)^4|Z=z\right] \leq U(Q)$.

(iii) $\mu_-(z,Q)$ *and* $\mu_+(z,Q)$ *are 3 times continuously differentiable, and* $|\mu_-^v(z,Q)| \leq U(Q)$ *and* $|\mu_+^v(z,Q)| \leq U(Q)$ *for* $v = 1,2,3$.

(iv) $\sigma_-^2(z,Q)$ *and* $\sigma_+^2(z,Q)$ *are continuous, and* $L(Q) \leq \sigma_-^2(z,Q) \leq U(Q)$ *and* $L(Q) \leq \sigma_+^2(z,Q) \leq U(Q)$.

In this setting, our parameter of interest is the average treatment effect (ATE) at the threshold,

$$\theta(Q) = \mu_+(0,Q) - \mu_-(0,Q) \ . \tag{3.6}$$

The above parameter is identified, as shown by Hahn et al. (2001b), using the distribution of the observed random variables by

$$\theta(P) = \lim_{z \to 0^+} \mu(z,P) - \lim_{z \to 0^-} \mu(z,P) \ , \tag{3.7}$$

where $\mu(z,P) = E_P[Y|Z=z]$. The hypotheses of interest can then be stated as

$$H_0 : P \in \mathbf{P}_0 \ \text{ versus } \ H_1 : P \in \mathbf{P}_1 = \mathbf{P} \setminus \mathbf{P}_0 \ , \tag{3.8}$$

where $\mathbf{P}_0 = \{P \in \mathbf{P} \mid \theta(P) = \theta_0\}$ is the subset of $\mathbf{P}$ for which the null hypothesis that the ATE at the threshold equals a pre-specified value of θ_0 holds.

Remark 3.2.1. To be concise, we focus on the ATE in the so-called sharp RDD (characterized by the treatment assignment rule in (3.2)). Our results in Section 3.4 will however follow with some manipulation for other parameters such as quantiles, and for

parameters in other designs such as the kink RDD in Card et al. (2015) or the fuzzy RDD.

■

3.3. Testing Problem

The testing problem we study is to ideally construct a finite sample test $\phi = \phi(X^{(n)})$ for (3.8). A requirement of the test is that it controls size, which is said to be level α whenever

$$\sup_{P \in \mathbf{P}_0} E_{P^n}[\phi] \leq \alpha \ , \tag{3.9}$$

where $\alpha \in (0, 1)$ is the chosen level of significance. Note that the above is a finite sample requirement, and to construct nontrivial tests that control size in finite samples may be too demanding. Alternatively, we also study the construction of a sequence of tests $\{\phi_n\}_{n=1}^{\infty}$ that are required to control limiting size, i.e.

$$\limsup_{n \to \infty} \sup_{P \in \mathbf{P}_0} E_{P^n}[\phi_n] \leq \alpha \ , \tag{3.10}$$

and are said to be uniformly asymptotically level α. As highlighted in Remark 3.4.3, this requirement is in contrast to the one for pointwise asymptotically valid tests where (3.10) is not required to hold uniformly across distributions in $\mathbf{P}_0$.

In our results, we show that under the commonly imposed setup described in the previous section, it is impossible to construct nontrivial tests that satisfy (3.9) or sequence of such asymptotically nontrivial tests that satisfy (3.10). We achieve this by illustrating that (3.8) has the property such that for any test ϕ the power against any alternative is

bounded above by its size, i.e.

$$\sup_{P\in\mathbf{P}_1} E_{P^n}[\phi] \leq \sup_{P\in\mathbf{P}_0} E_{P^n}[\phi] \ . \tag{3.11}$$

To prove this claim, we rely on an insightful result from Romano (2004) restated in the following lemma for clarity, where

$$\tau(P,P') = \sup_{\{g:|g|\leq 1\}} \left| \int g dP - \int g dP' \right| \tag{3.12}$$

denotes the total variation metric between any two distributions P and P'. This lemma additionally formalizes the concept of what we mean by $\mathbf{P}$ (and hence $\mathbf{Q}$) being large in some sense.

Lemma 3.3.1. *Let $n \geq 1$ and ϕ be any test of $\mathbf{P}_0$ versus $\mathbf{P}_1$ in (3.8). If for every $P \in \mathbf{P}_1$ there exists a sequence $\{P_k\}_{k=1}^{\infty}$ in $\mathbf{P}_0$ such that $\tau(P,P_k) \to 0$ as $k \to \infty$, then*

$$\sup_{P\in\mathbf{P}_1} E_{P^n}[\phi] \leq \sup_{P\in\mathbf{P}_0} E_{P^n}[\phi] \ . \tag{3.13}$$

3.4. Main Results

3.4.1. Testability in the Basic Setup

In the following theorem we establish that when $\mathbf{Q}$ is defined as in (3.5), any test for (3.8) will have power against any alternative bounded above by its size.

Theorem 3.4.1. *Let $n \geq 1$, $\mathbf{Q}$ be defined as in* (3.5)*, $\mathbf{P}$ be as in* (3.4) *and $\mathbf{P}_0$ and $\mathbf{P}_1$ be as in* (3.8)*. Then any test ϕ satisfies*

$$\sup_{P \in \mathbf{P}_1} E_{P^n}[\phi] \leq \sup_{P \in \mathbf{P}_0} E_{P^n}[\phi] \ . \tag{3.14}$$

Proof. Fix $P \in \mathbf{P}_1$ and take any strictly positive sequence $\{\epsilon_k\}_{k=1}^{\infty}$ such that $\epsilon_k \to 0$ as $k \to \infty$. Since (3.4) implies that $P = QM^{-1}$ for some $Q \in \mathbf{Q}$, it then follows from Assumption 3.2.1 (i) that for every k there exists a Borel set B_k in $\mathcal{X}$,

$$B_k = \{(y, z) \in \mathcal{X} : z \in (-\tilde{\epsilon}_k, \tilde{\epsilon}_k)\} \ , \tag{3.15}$$

where $\tilde{\epsilon}_k > 0$, such that $0 < P(B_k) \leq \epsilon_k$. Take next any $P' \in \mathbf{P}_0$ that has the same density on Z as P. We may then construct the sequence $\{P_k\}_{k=1}^{\infty}$ such that for every Borel subset B of $\mathcal{X}$ let

$$P_k(B) := P(B \cap B_k^c) + P'(B \cap B_k) \ , \tag{3.16}$$

where B_k^c denotes the complement of B_k. One can verify that for every k that P_k is a well defined distribution.

Next, we show that $\{P_k\}_{k=1}^{\infty}$ is in $\mathbf{P}_0$, i.e. for every k there exists $Q_k \in \mathbf{Q}$ such that $\theta(Q_k) = \theta_0$ and $P_k = Q_k M^{-1}$. To construct this Q_k, first note that $P = QM^{-1}$ and $P' = Q'M^{-1}$ for some $Q \in \mathbf{Q}$ and $Q' \in \mathbf{Q}$ with $\theta(Q') = \theta_0$. Then for every Borel subset $\tilde{B}$ of $\mathcal{W}$ let

$$Q_k(\tilde{B}) := Q(\tilde{B} \cap \tilde{B}_k^c) + Q'(\tilde{B} \cap \tilde{B}_k) \ , \tag{3.17}$$

where

$$\tilde{B}_k = M^{-1}(B_k) = \{(y_0, y_1, z) \in \mathcal{W} : z \in (-\tilde{\epsilon}_k, \tilde{\epsilon}_k)\} \ , \tag{3.18}$$

and $\tilde{B}_k^c$ denotes the complement of $\tilde{B}_k$, which in this case is just $M^{-1}(B_k^c)$. Analogous to P_k, it follows that Q_k is a well defined distribution. To show that $Q_k \in \mathbf{Q}$, first note (3.17) ensures that $Q_k(A) = Q'(A)$ for every Borel subset A of $\mathcal{W}$ that satisfies $A \subseteq \tilde{B}_k$. This implies that the density and all the conditional on $Z = z$ quantities in Assumption 3.2.1 are equal for Q_k and Q' for all $z \in (-\tilde{\epsilon}_k, \tilde{\epsilon}_k)$. In turn, it follows that Q_k satisfies Assumption 3.2.1 by taking $\kappa(Q_k) = \min\{\kappa(Q'), \tilde{\epsilon}_k\}$, $L(Q_k) = L(Q')$, and $U(Q_k) = U(Q')$. Further, by the same argument, it follows that $\theta(Q_k) = \theta_0$ as $\theta(Q') = \theta_0$. Finally, given that for every $B \subseteq \mathcal{X}$ and $\tilde{B} = M^{-1}(B)$ we have $\tilde{B} \cap \tilde{B}_k^c = M^{-1}(B \cap B_k^c)$ and $\tilde{B} \cap \tilde{B}_k = M^{-1}(B \cap B_k)$, we can establish $P_k = Q_k M^{-1}$ from (3.16) and (3.17).

To conclude, we show that the total variation distance between P and P_k goes to 0 as $k \to \infty$,

$$\begin{aligned} \tau(P, P_k) &= \sup_{\{g:|g|\leq 1\}} \left| \int g dP - \int g dP_k \right| = \sup_{\{g:|g|\leq 1\}} \left| \int_{B_k} g dP - \int_{B_k} g dP' \right| \\ &\leq \left| \int_{B_k} dP \right| + \left| \int_{B_k} dP' \right| \leq 2\epsilon_k \to 0 \ , \end{aligned} \tag{3.19}$$

where the second and fourth relations follow from (3.16) and (3.15) respectively, along with noting that P and P' have the same density for Z. Since $P \in \mathbf{P}_1$ was chosen arbitrarily, we can then invoke Lemma 3.3.1 to conclude the proof. ■

Remark 3.4.1. In invoking Lemma 3.3.1 to prove Theorem 3.4.1, for any $P \in \mathbf{P}_1$ we construct a sequence $\{P_k\}_{k=1}^{\infty}$ in $\mathbf{P}_0$, such that for every k there exists a Borel set in

$\mathcal{X}$ with positive probability under P_k where P and P_k differ, and are otherwise equal on the complement of this set. Letting the probability of this set vanish with k implies that $\tau(P, P_k) \to 0$ as $k \to \infty$. Further, since Assumption 3.2.1 only requires conditions local to zero that are pointwise in nature, we formally show in the proof that this ensures that our construction $\{P_k\}_{k=1}^{\infty}$ falls in $\mathbf{P}_0$. Note that our construction is not unique and that multiple others are possible. ■

Remark 3.4.2. It is important to emphasize that Theorem 3.4.1 is not a criticism of a specific test but holds for any choice of test. Furthermore, it is a statement on the finite sample property of any test, but with important asymptotic implications. To be specific, for any sequence of tests $\{\phi_n\}_{n=1}^{\infty}$ with nontrivial limiting power, it directly follows from (3.14) that

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} E_{P^n}[\phi_n] > \alpha \ . \tag{3.20}$$

This additionally implies that if the sequence of tests is pointwise consistent in power, i.e. pointwise power converges to one, then limiting size is in fact equal to one. ■

Remark 3.4.3. Currently used tests are shown to be only pointwise asymptotically valid, i.e.

$$\limsup_{n\to\infty} E_{P^n}[\phi_n] \leq \alpha \text{ for all } P \in \mathbf{P}_0 \ , \tag{3.21}$$

which does not say anything about whether this sequence of tests $\{\phi_n\}_{n=1}^{\infty}$ approximates (3.9) for large enough n. To be specific, it is possible that for every $n \geq 1$ there exists

$P \in \mathbf{P}_0$ such that

$$E_{P^n}[\phi_n] >> \alpha \ . \tag{3.22}$$

■

3.4.2. Uniformly Valid Test under Stronger Assumptions

In this section, we ask under what alternative assumptions we can construct a uniformly asymptotically valid test. We consider, in particular, a natural strengthening of Assumption 3.2.1 leading to the following alternative definition of the set of possible distributions,

$$\mathbf{Q} = \{Q \in \mathbf{Q}_{\mathcal{W}} : Q \text{ satisfies Assumption 3.4.1}\} \ , \tag{3.23}$$

where as before $\mathbf{Q}_{\mathcal{W}}$ denotes the set of all Borel probability measures on $\mathcal{W}$ that have a density on Z with respect to the Lebesgue measure, and Assumption 3.4.1 is stated below. Note that if Q satisfies Assumption 3.4.1 then it satisfies Assumption 3.2.1, and hence the definition of $\mathbf{Q}$ in (3.23) generates a smaller set of distributions than the definition of $\mathbf{Q}$ in (C.3).

Assumption 3.4.1. *Let Q be such that it satisfies Assumption 3.2.1 with $\kappa(Q) = \tilde{\kappa}$, $L(Q) = \tilde{L}$ and $U(Q) = \tilde{U}$, where $\tilde{\kappa} > 0$ and $\tilde{U} > \tilde{L} > 0$ are real numbers that do not depend on Q.*

We next briefly describe a simple version of the Calonico et al. (2014a) test (referred to as CCT hereafter), which is demonstrated to satisfy (3.10) under this smaller set of

distributions. For the null hypothesis in (3.8), the CCT test statistic is

$$T_n^{CCT}(X^{(n)}) = \frac{\hat{\theta}_n - \theta_0}{\hat{S}_n} \ , \tag{3.24}$$

where $\hat{\theta}_n$ is a bias corrected local linear estimator of $\theta(P)$, and $\hat{S}_n$ is a plug-in estimator of a novel standard error formula that accounts for the variance of the bias estimate. The bias is estimated using a local quadratic estimator. Furthermore, for all estimates, we use the triangular kernel and a deterministic sequence of bandwidth choices denoted by h_n. Then, the CCT level α test is

$$\phi_n^{CCT}(X^{(n)}) = \begin{cases} 1 \text{ if } |T_n^{CCT}(X^{(n)})| > z_{1-\alpha/2} \\ \\ 0 \text{ otherwise} \end{cases} \ , \tag{3.25}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-quantile of the standard normal distribution.

The following theorem demonstrates that under the alternative definition of $\mathbf{Q}$ in (3.23), the test statistic in (3.24) for (3.8) asymptotically converges uniformly in $\mathbf{P}_0$ to the standard normal distribution. It then directly follows that the test in (3.25) is uniformly asymptotically level α, and, in fact, has limiting size equal to α.

Theorem 3.4.2. *Let* $\mathbf{Q}$ *be defined as in* (3.23), $\mathbf{P}$ *be as in* (3.4) *and* $\mathbf{P}_0$ *and* $\mathbf{P}_1$ *be as in* (3.8). *If* $nh_n \to \infty$, $h_n \to 0$ *and* $nh_n^7 \to 0$, *then the CCT test statistic from* (3.24) *satisfies*

$$T_n^{CCT}(X^{(n)}) = \frac{\hat{\theta}_n - \theta_0}{\hat{S}_n} \xrightarrow{d} \mathcal{N}(0,1) \tag{3.26}$$

as $n \to \infty$, where $X^{(n)}$ are i.i.d P_n and P_n is any sequence of distributions such that $P_n \in \mathbf{P}_0$ for all $n \geq 1$. This in turn implies that $\{\phi_n^{CCT}\}_{n=1}^{\infty}$ in (3.25) is uniformly asymptotically level α, and, in fact, has limiting size equal to α, i.e.

$$\limsup_{n\to\infty} \sup_{P\in\mathbf{P}_0} E_{P^n}[\phi_n^{CCT}] = \alpha \ . \tag{3.27}$$

The proof of the above essentially requires slightly altering the proof of the pointwise result in Calonico et al. (2014a) to any sequence of distributions P_n such that $P_n \in \mathbf{P}_0$ for all $n \geq 1$. For completeness, we provide a proof in Appendix C.

Remark 3.4.4. Note that when $\mathbf{Q}$ is defined as in (3.23) the arguments used to prove Theorem 3.4.1 do not go through. In particular, the constructed sequence $\{P_k\}_{k=1}^{\infty}$ in (3.16) will not fall in $\mathbf{P}$, as the corresponding $\{Q_k\}_{k=1}^{\infty}$ in (3.17) will not fall in $\mathbf{Q}$. To see why, note that for large enough k we have $\kappa(Q_k) < \tilde{\kappa}$, and either $\mu_-(z, Q_k)$ or $\mu_+(z, Q_k)$ is discontinuous at either $z = -\kappa(Q_k)$ or $z = \kappa(Q_k)$. This implies Q_k will not satisfy Assumption 3.4.1 as $\mu_-(z, Q_k)$ or $\mu_+(z, Q_k)$ will not be continuous for all $z \in (-\tilde{\kappa}, \tilde{\kappa})$,. Intuitively, Assumption 3.4.1 excludes extreme sequences such as $\{Q_k\}_{k=1}^{\infty}$ for which nonparametric tools work poorly to give a uniform limit result. For recent additional results on uniform testing in RDD, see Armstrong and Kolesar (2016) and Calonico et al. (2016). ■

Bibliography

Almond, D., Doyle Jr, J. J., Kowalski, A. and Williams, H. (2010). Estimating marginal returns to medical care: Evidence from at-risk newborns. *The Quarterly Journal of Economics*, **125** 591–634.

Andrews, D. W. and Guggenberger, P. (2009a). Incorrect asymptotic size of subsampling procedures based on post-consistent model selection estimators. *Journal of Econometrics*, **152** 19 – 27.

Andrews, D. W. and Guggenberger, P. (2009b). Validity of subsampling and plug-in asymptotic inference for parameters defined by moment inequalities. *Econometric Theory*, **25** 669–709.

Angrist, J. D. and Imbens, G. W. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American statistical Association*, **90** 431–442.

Armstrong, T. and Kolesar, M. (2016). Optimal inference in a class of regression models. Manuscript.

Bahadur, R. R. and Savage, L. J. (1956). The nonexistence of certain statistical procedures in nonparametric problems. *Ann. Math. Statist.*, **27** 1115–1122.

Ben-Akiva, M. and Boccara, B. (1995). Discrete choice models with latent choice sets. *International journal of Research in Marketing*, **12** 9–24.

Bhattacharya, P. (1974). Convergence of sample paths of normalized sums of induced order statistics. *The Annals of Statistics* 1034–1039.

Bloom, H. S. and Weiland, C. (2015). Quantifying variation in head start effects on young children's cognitive and socio-emotional skills using data from the national head start impact study.

Blundell, R., Gosling, A., Ichimura, H. and Meghir, C. (2007). Changes in the distribution of male and female wages accounting for employment composition using bounds. *Econometrica*, **75** 323–363.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Bruhn, M. and McKenzie, D. (2008). In pursuit of balance: Randomization in practice in development field experiments. *World Bank Policy Research Working Paper*, **4752**.

Bugni, F. A. (2016). Comparison of inferential methods in partially identified models in terms of error in coverage probability. *Econometric Theory*, **32** 187242.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2016). Inference under covariate-adaptive randomization. CeMMAP working paper CWP45/15.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2017a). Inference under covariate-adaptive randomization. *Journal of the American Statistical Association*.

Bugni, F. A., Canay, I. A. and Shaikh, A. M. (2017b). Inference under covariate adaptive randomization with multiple treatments.

Calonico, S., Cattaneo, M. D. and Farrell, M. H. (2016). Coverage error optimal confidence intervals for regression discontinuity designs. Manuscript.

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

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014b). Robust nonparametric confidence intervals for regression-discontinuity designs. *Econometrica*, **82**.

Calonico, S., Cattaneo, M. D. and Titiunik, R. (2014c). Supplement to "Robust nonparametric confidence intervals for regression-discontinuity designs". *Econometrica Supplement Material*, **82**.

Canay, I. A. and Kamat, V. (2017). Approximate permutation tests and induced order statistics in the regression discontinuity design. *The Review of Economic Studies*.

Canay, I. A., Romano, J. P. and Shaikh, A. M. (2017). Randomization tests under an approximate symmetry assumption. *Econometrica*, **85** 1013–1030.

Canay, I. A., Santos, A. and Shaikh, A. M. (2013). On the testability of identification in some non-parametric models with endogeneity. *Econometrica*, **81** 2535–2559.

Canay, I. A. and Shaikh, A. M. (2017). *Practical and Theoretical Advances in Inference for Partially Identified Models*, vol. 2 of *Econometric Society Monographs*. Cambridge University Press, 271306.

Card, D., Lee, D. S., Pei, Z. and Weber, A. (2015). Inference on causal effects in a generalized regression kink design. *Econometrica*, **83** 2453–2483.

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.

Charnes, A. and Cooper, W. W. (1962). Programming with linear fractional functionals. *Naval Research Logistics (NRL)*, **9** 181–186.

Chung, E. and Romano, J. P. (2013). Exact and asymptotically robust permutation tests. *The Annals of Statistics*, **41** 484–507.

David, H. and Galambos, J. (1974). The asymptotic theory of concomitants of order statistics. *Journal of Applied Probability* 762–770.

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica*, **65** pp. 1365–1387.

Dufour, J.-M. (2003). Identification, weak instruments, and statistical inference in econometrics. *Canadian Journal of Economics/Revue canadienne d'conomique*, **36** 767–808.

Frandsen, B. R., Frlich, M. and Melly, B. (2012). Quantile treatment effects in the regression discontinuity design. *Journal of Econometrics*, **168** 382 – 395.

Ganong, P. and Jäger, S. (2015). A permutation test for the regression kink design. Tech. rep., Working paper.

Gerard, F., Rokkanen, M. and Rothe, C. (2016). Bounds on treatment effects in regression discontinuity designs with a manipulated running variable, with an application to unemployment insurance in brazil. Working paper.

Guggenberger, P. (2010a). The impact of a Hausman pretest on the asymptotic size of a hypothesis test. *Econometric Theory*, **26** 369–382.

Guggenberger, P. (2010b). The impact of a Hausman pretest on the size of a hypothesis test: The panel data case. *Journal of Econometrics*, **156** 337 – 343.

Hahn, J., Todd, P. and Klaauw, W. V. d. (2001a). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, **69** pp. 201–209. URL `http://www.jstor.org/stable/2692190`.

Hahn, J., Todd, P. and Van der Klaauw, W. (2001b). Identification and estimation of treatment effects with a regression-discontinuity design. *Econometrica*, **69** 201–209.

Hajek, J., Sidak, Z. and Sen, P. K. (1999). *Theory of rank tests*. 2nd ed. Academic press.

Heckman, J. J. and Honore, B. E. (1990). The empirical content of the roy model. *Econometrica: Journal of the Econometric Society* 1121–1149.

Heckman, J. J. and Pinto, R. (2017). Unordered monotonicity. Tech. rep., National Bureau of Economic Research.

Heckman, J. J., Smith, J. and Clements, N. (1997). Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies*, **64** 487–535.

Heckman, J. J., Urzua, S. and Vytlacil, E. (2006). Understanding instrumental variables in models with essential heterogeneity. *The Review of Economics and Statistics*, **88** 389–432.

Heckman, J. J., Urzua, S. and Vytlacil, E. (2008). Instrumental variables in models with multiple outcomes: The general unordered case. *Annales d'Economie et de Statistique* 151–174.

Heckman, J. J. and Vytlacil, E. J. (2007). Econometric evaluation of social programs, part ii: Using the marginal treatment effect to organize alternative econometric estimators to evaluate social programs, and to forecast their effects in new environments. *Handbook of econometrics*, **6** 4875–5143.

Hoeffding, W. (1952). The large-sample power of tests based on permutations of observations. *The Annals of Mathematical Statistics*, **23** pp. 169–192. URL `http:`

//www.jstor.org/stable/2236445.

Hull, P. (2015). Isolateing: Identifying counterfactual-specific treatment effects with cross-stratum comparisons.

Imbens, G. and Kalyanaraman, K. (2012a). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, **79** 933–959.

Imbens, G. and Kalyanaraman, K. (2012b). Optimal bandwidth choice for the regression discontinuity estimator. *The Review of Economic Studies*, **79** 933–959.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, **62** 467–475.

Imbens, G. W. and Lemieux, T. (2008a). Regression discontinuity designs: A guide to practice. *Journal of econometrics*, **142** 615–635.

Imbens, G. W. and Lemieux, T. (2008b). Regression discontinuity designs: A guide to practice. *Journal of Econometrics*, **142** 615–635.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, **72** 1845–1857.

Kaido, H., Molinari, F. and Stoye, J. (2016). Confidence intervals for projections of partially identified parameters. *arXiv preprint arXiv:1601.00934*.

Kamat, V. (2017). On nonparametric inference in the regression discontinuity design. *Econometric Theory* 1–10.

Kirkeboen, L. J., Leuven, E. and Mogstad, M. (2016). Field of study, earnings, and self-selection. *The Quarterly Journal of Economics*, **131** 1057–1111.

Kline, P. and Tartari, M. (2016). Bounding the labor supply responses to a randomized welfare experiment: A revealed preference approach. *The American Economic*

Review, **106** 971–1013.

Kline, P. and Walters, C. R. (2016). Evaluating public programs with close substitutes: The case of head start. *The Quarterly Journal of Economics*, **131** 1795–1848.

Lee, D. S. (2008). Randomized experiments from non-random selection in u.s. house elections. *Journal of Econometrics*, **142** 675 – 697. The regression discontinuity design: Theory and applications, URL `http://www.sciencedirect.com/science/article/pii/S0304407607001121`.

Lee, D. S. and Lemieux, T. (2010a). Regression discontinuity designs in economics. *Journal of Economic Literature*, **48** 281–355.

Lee, D. S. and Lemieux, T. (2010b). Regression discontinuity designs in economics. *Journal of Economic Literature*, **48** 281–355.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econometric Theory* 21–59.

Leeb, H. and Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, **24** 338–376.

Lehmann, E. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. 3rd ed. Springer, New York.

Lehmann, E. L. and Loh, W.-Y. (1990). Pointwise versus uniform robustness of some large-sample tests and confidence intervals. *Scandinavian Journal of Statistics*, **17** pp. 177–187.

Manski, C. F. (1996). Learning about treatment effects from experiments with random assignment of treatments. *Journal of Human Resources* 709–733.

Manski, C. F. (1997a). The mixing problem in programme evaluation. *The Review of Economic Studies*, **64** 537–553.

Manski, C. F. (1997b). Monotone treatment response. *Econometrica: Journal of the Econometric Society* 1311–1334.

Manski, C. F. (2003). *Partial identification of probability distributions*. Springer Science & Business Media.

Manski, C. F. (2007). Partial identification of counterfactual choice probabilities. *International Economic Review*, **48** 1393–1410.

Manski, C. F. (2014). Identification of income–leisure preferences and evaluation of income tax policy. *Quantitative Economics*, **5** 145–174.

Manski, C. F. and Pepper, J. V. (2000). Monotone instrumental variables: With an application to the returns to schooling. *Econometrica*, **68** 997–1010.

Manski, C. F. and Pepper, J. V. (2009). More on monotone instrumental variables. *The Econometrics Journal*, **12**.

Marschak, J. et al. (1959). Binary choice constraints on random utility indicators. Tech. rep., Cowles Foundation for Research in Economics, Yale University.

McCrary, J. (2008a). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, **142** 698 – 714. The regression discontinuity design: Theory and applications, URL `http://www.sciencedirect.com/science/article/pii/S0304407607001133`.

McCrary, J. (2008b). Manipulation of the running variable in the regression discontinuity design: A density test. *Journal of Econometrics*, **142** 698 – 714.

Mikusheva, A. (2007). Uniform inference in autoregressive models. *Econometrica*, **75** 1411–1452.

Mikusheva, A. (2010). Robust confidence sets in the presence of weak instruments. *Journal of Econometrics*, **157** 236 – 247.

Mikusheva, A. (2012). One-dimensional inference in autoregressive models with the potential presence of a unit root. *Econometrica*, **80** 173–212.

Mogstad, M., Santos, A. and Torgovitsky, A. (2017). Using instrumental variables for inference about policy relevant treatment effects. Tech. rep., National Bureau of Economic Research.

Mourifie, I., Henry, M. and Meango, R. (2015). Sharp bounds for the roy model.

Müller, U. K. (2008). The impossibility of consistent discrimination between i(0) and i(1) processes. *Econometric Theory*, **24** 616–630.

Otsu, T., Xu, K.-L. and Matsushita, Y. (2015). Empirical likelihood for regression discontinuity design. *Journal of Econometrics*, **186** 94 – 112.

Politis, D. N. and Romano, J. P. (1994). Large sample confidence regions based on subsamples under minimal assumptions. *The Annals of Statistics* 2031–2050.

Puma, M., Bell, S., Cook, R., Heid, C., Shapiro, G., Broene, P., Jenkins, F., Fletcher, P., Quinn, L., Friedman, J. et al. (2010). Head start impact study. final report. *Administration for Children & Families*.

Romano, J. P. (1989). Bootstrap and randomization tests of some nonparametric hypotheses. *The Annals of Statistics*, **17** 141–159. URL `http://dx.doi.org/10.1214/aos/1176347007`.

Romano, J. P. (1990). On the behavior of randomization tests without a group invariance assumption. *Journal of the American Statistical Association*, **85** pp. 686–692. URL `http://www.jstor.org/stable/2290003`.

Romano, J. P. (2004). On non-parametric testing, the uniform behaviour of the t-test, and related problems. *Scandinavian Journal of Statistics*, **31** 567–584.

Romano, J. P. and Shaikh, A. M. (2008a). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, **138** 2786–2807.

Romano, J. P. and Shaikh, A. M. (2008b). Inference for identifiable parameters in partially identified econometric models. *Journal of Statistical Planning and Inference*, **138** 2786–2807.

Romano, J. P., Shaikh, A. M. et al. (2012). On the uniform asymptotic validity of subsampling and the bootstrap. *The Annals of Statistics*, **40** 2798–2822.

Sales, A. and Hansen, B. B. (2015). Limitless regression discontinuity. *arXiv preprint arXiv:1403.5478*.

Sekhon, J. S. and Titiunik, R. (2016). On interpreting the regression discontinuity design as a local experiment. Manuscript.

Shen, S. and Zhang, X. (2016). Distributional tests for regression discontinuity: Theory and empirical examples. *Review of Economics and Statistics*. Forthcoming.

Stopher, P. R. (1980). Captivity and choice in travel-behavior models. *Transportation engineering journal of the American Society of Civil Engineers*, **106** 427–435.

Stoye, J. (2010). Partial identification of spread parameters. *Quantitative Economics*, **1** 323–357.

Torgovitsky, A. (2016). Nonparametric inference on state dependence with applications to employment dynamics.

Torgovitsky, A. (2017). Partial identification by extending subdistributions.

van der Vaart, A. W. (1998a). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. (1998b). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Walters, C. R. (2015). Inputs in the production of early childhood human capital: Evidence from head start. *American Economic Journal: Applied Economics*, **7** 76–102.

Williams, H. and Ortúzar, J. d. D. (1982). Behavioural theories of dispersion and the mis-specification of travel demand models. *Transportation Research Part B: Methodological*, **16** 167–219.

APPENDIX A

Appendix to Chapter 1

A.0.1. Restatement of Test Statistic

In this appendix, I describe how the test statistic in (1.31) can be restated as a solution to a linear program. This restatement in particular follows similar ones stated in Mogstad et al. (2017) and Torgovitsky (2016). To this end, begin by noting that test statistic can be explicitly be written as

$$TS_n(\theta_0) = \sqrt{G} \min_{\{Q(w)\}_{w\in\mathcal{W}}} \sum_{x\in\mathcal{X}} |\hat{m}_{\text{dat},x}(Q)| + \sum_{s\in\mathcal{S}_2} |\hat{m}_s(Q)| \ ,$$

subject to the following constraints:

(i) $\sum_{w\in\mathcal{W}} a_{\text{num}}(w)\cdot Q(w) = \theta_0 \cdot \sum_{w\in\mathcal{W}} a_{\text{den}}(w)\cdot Q(w)$.
(ii) $0 \leq Q(w) \leq 1$ for every $w \in \mathcal{W}$.
(iii) $\sum_{w\in\mathcal{W}} Q(w) = 1$.
(iv) $\sum_{w\in\mathcal{W}} a_s(w)\cdot Q(w) \leq b_s$ for every $s \in \mathcal{S}_1$.

By introducing two additional variables for each data restriction and stochastic restriction, it can be shown that the above program can be re-written as a linear program. More specifically, the above optimization problem is equivalent to

$$TS_n(\theta_0) = \sqrt{G} \min_{\substack{\{Q(w)\}_{w\in\mathcal{W}},\\ \{\mu_1^+(x)\}_{x\in\mathcal{X}},\{\mu_1^-(x)\}_{x\in\mathcal{X}},\\ \{\mu_2^+(s)\}_{s\in\mathcal{S}_2},\{\mu_2^-(s)\}_{s\in\mathcal{S}_2}}} \sum_{x\in\mathcal{X}} (\mu_1^+(x)+\mu_1^-(x)) + \sum_{s\in\mathcal{S}_1} (\mu_2^+(s)+\mu_2^-(s)) \ ,$$

subject to the following constraints:

(i) $\sum_{w \in \mathcal{W}} a_{\text{num}}(w) \cdot Q(w) = \theta_0 \cdot \sum_{w \in \mathcal{W}} a_{\text{den}}(w) \cdot Q(w)$.

(ii) $0 \leq Q(w) \leq 1$ for every $w \in \mathcal{W}$.

(iii) $\sum_{w \in \mathcal{W}} Q(w) = 1$.

(iv) $\sum_{w \in \mathcal{W}} a_s(w) \cdot Q(w) \leq b_s$ for every $s \in \mathcal{S}_1$.

(v) $\mu_1^+(x) \geq 0$ and $\mu_1^-(x) \geq 0$ for every $x \in \mathcal{X}$.

(vi) $\mu_2^+(s) \geq 0$ and $\mu_2^-(s) \geq 0$ for every $s \in \mathcal{S}_2$.

(vii) $\hat{m}_{\text{dat},x}(Q) = \mu_1^+(x) - \mu_1^-(x)$ for every $x \in \mathcal{X}$.

(viii) $\hat{m}_s(Q) = \mu_2^+(s) - \mu_2^-(s)$ for every $s \in \mathcal{S}_2$.

Since, as noted in Section 1.5, both the moment conditions $\hat{m}_{\text{dat},x}(Q)$ and $\hat{m}_s(Q)$ are linear in Q, the above stated problem is a linear program. For a discussion of such a restatement, see (Boyd and Vandenberghe, 2004, Page 294).

A.0.2. Restrictions Imposed by Additional Identifying Assumptions

Lemma A.0.1. *Assumption Roy imposes restrictions on Q that satisfy Assumption 1.4.2.*

Proof: In order to see the restriction imposed by Assumption Roy, note first that this assumption can be re-written as

$$d_{U,\{d,d'\}} = d \implies Y(d') \leq Y(d)$$

for every $d, d' \in \mathcal{D}$. Then, denoting by

$$\mathcal{U}_{\{d,d'\}} = \{u \in \mathcal{U} : d_{u,\{d,d'\}} = d\}$$

note the above statement imposes that

$$\text{Prob}_Q[Y^\dagger(d) = 0, Y^\dagger(d') = 1, U \in \mathcal{U}_{\{d,d'\}}] = 0$$

for every $d, d' \in \mathcal{D}$. Equivalently, using the notation introduced in Lemma 1.4.1, this can re-written as a linear restriction on Q in the form of Assumption 1.4.2 as

$$\sum_{\bar{y} \in \mathcal{Y}_{\{d,d'\}}, u \in \mathcal{U}_{\{d,d'\}}, \bar{c} \in \mathcal{C}^2, z \in \{0,1\}, \bar{z} \in \{0,1\}^2} Q(\bar{y}, u, \bar{c}, z, \bar{z}) = 0$$

for every $d, d' \in \mathcal{D}$, where

$$\mathcal{Y}_{\{d,d'\}} = \{\bar{y} \in \{0,1\}^3 : y(d) = 0, y(d') = 1\} \ .$$

■

Lemma A.0.2. *Assumption MTR imposes restrictions on Q that satisfy Assumption 1.4.2.*

Proof: In order to see the restriction imposed by Assumption MTR, note first that this assumption imposes that

$$\text{Prob}_Q[Y^\dagger(0) = 1, Y^\dagger(d) = 0] = 0 \ ,$$

for each $d \in \{1, 2\}$. Equivalently, using the notation introduced in Lemma 1.4.1, this can then be equivalently re-written as a linear restriction on Q in the form of Assumption

1.4.2 as

$$\sum_{\bar{y}\in\mathcal{Y}_d,u\in\mathcal{U},\bar{c}\in\mathcal{C}^2,z\in\{0,1\},\bar{z}\in\{0,1\}^2} Q(\bar{y},u,\bar{c},z,\bar{z}) = 0$$

for each $d \in \{1,2\}$, where

$$\mathcal{Y}_d = \{\bar{y} \in \{0,1\}^3 : y(0) = 1, y(d) = 0\} \ .$$

■

Lemma A.0.3. *Assumption SLI imposes restrictions on Q that satisfy Assumption 1.4.2.*

PROOF: In order to see the restriction imposed by Assumption SLI, note first that this independence assumption imposes that

$$\text{Prob}_Q[Y^\dagger(0) = y, HC = hc, CS = cs] = \text{Prob}_Q[Y^\dagger(0) = y] \cdot \text{Prob}_Q[HC = hc, CS = cs]$$

for $y, hc, cs \in \{0,1\}$. Equivalently, using the notation introduced in Lemma 1.4.1, this can then be equivalently re-written as a linear restriction on Q in the form of Assumption 1.4.2 as

$$\sum_{\bar{y}\in\mathcal{Y}_y,u\in\mathcal{U},\bar{c}\in\mathcal{C}^2,z\in\{0,1\}} Q(\bar{y},u,\bar{c},z,\bar{z})$$
$$- \sum_{y\in\{0,1\},d\in\mathcal{D},z\in\{0,1\}} P(y,d,z,\bar{z}) \cdot \sum_{\bar{y}\in\mathcal{Y}_y,u\in\mathcal{U},\bar{c}\in\mathcal{C}^2,z\in\{0,1\},\bar{z}\in\{0,1\}^2} Q(\bar{y},u,\bar{c},z,\bar{z}) = 0$$

for each $y \in \{0,1\}$ and $\bar{z} \in \{0,1\}^2$, where

$$\mathcal{Y}_y = \{\bar{y} \in \{0,1\}^3 : y(0) = y\} \ .$$

■

Lemma A.0.4. *Assumption UA imposes a restriction on Q that satisfies Assumption 1.4.2.*

PROOF: In order to see the restriction imposed by Assumption UA, note first that this assumption can be written as

$$\text{Prob}_Q[(C(0), C(1)) \in \mathcal{C}_{\text{UA}}] = 1 \ ,$$

where

$$\mathcal{C}_{\text{UA}} = \{(\{0\},\{0,2\}),(\{0,2\},\{0,2\}),(\{0,1\},\{0,1,2\}),(\{0,1,2\},\{0,1,2\})\}$$

is the set of all combinations of choice with and without offer such that the choice set with an offer is the same as that without an offer except for the inclusion of or the lack of a Head Start preschool. Equivalently, this can be re-written as a linear restriction on Q in the form of Assumption 1.4.2 as

$$\sum_{w \in \mathcal{W}_{\text{UA}}} Q(w) = 1 \ ,$$

where $\mathcal{W}_{\text{UA}} = \{w \in \mathcal{W} : (c(0), c(1)) \in \mathcal{C}_{\text{UA}}\}$. ■

A.0.3. Data and Variable Construction

The raw data used from the HSIS in this paper is restricted, but access can be acquired by submitting applications to Research Connections at

`http://www.researchconnections.org/childcare/resources/19525` .

In this appendix, I briefly describe how the raw data was transformed to the final sample used in the empirical results in the paper, which closely followed the publicly available code used to construct the final sample in Kline and Walters (2016). I organize this description in the following steps which were taken separately for each age cohort:

Step 1: I merged all the various data files provided by Research Connections for the HSIS and dropped observations with missing Head Start center IDs, where this center corresponded to that from which the child was sampled. I then made edits to this raw sample as described below.

Step 2: I classified a variable for the Head Start HighScope curriculum and class size ratio covariate for each child from a given sampled Head Start center. This classification was required as in practice children from a given Head Start center could have attended different Head Start preschools. The covariate value from the modal attended Head Start preschool was used as the Head Start covariate value for all the children from that center.

Step 3: I classified the selected treatment intro the three categories used in the paper using the focal care arrangement variable provided by the HSIS data set.

Step 4: All observations where any of the variables used in the analysis were missing were dropped.

Step 5: Test score outcomes were then standardized using non-missing baseline test scores of the final sample. Moreover, the class size variable was transformed into a binary variable of high and low class size ratio, where high was taken to be above the median value across centers.

APPENDIX B

Appendix to Chapter 2

B.0.1. Proof of Theorem 2.4.1

First, note that the joint distribution of the induced order statistics

$$W^-_{n,[q]}, \ldots, W^-_{n,[1]}, W^+_{n,[1]}, \ldots, W^+_{n,[q]}$$

are conditionally independent given $(Z_1, \ldots, Z_n)$, with conditional cdfs

$$H(w|Z^-_{n,(q)}), \ldots, H(w|Z^-_{n,(1)}), H(w|Z^+_{n,(1)}), \ldots, H(w|Z^+_{n,(q)}) \ .$$

A proof of this result can be found in Bhattacharya (1974, Lemma 1). Now let $\mathcal{A} = \sigma(Z_1, \ldots, Z_n)$ be the sigma algebra generated by $(Z_1, \ldots, Z_n)$. It follows that

$$\begin{aligned} \Pr\left\{\bigcap_{j=1}^{q}\{W^-_{n,[j]} \le w^-_j\}\bigcap_{j=1}^{q}\{W^+_{n,[j]} \le w^+_j\}\right\} &= E\left[\Pr\left\{\bigcap_{j=1}^{q}\{W^-_{n,[j]} \le w^-_j\}\bigcap_{j=1}^{q}\{W^+_{n,[j]} \le w^+_j\}|\mathcal{A}\right\}\right] \\ &= E\left[\Pi^q_{j=1}H(w^-_j|Z^-_{n,(j)}) \cdot \Pi^q_{j=1}H(w^+_j|Z^+_{n,(j)})\right] \ . \end{aligned}$$

The first equality follow from the law of iterated expectations and the last equality follows from the conditional independence of the induced order statistics.

Let $f_{n,(q^-,\ldots,q^+)}(z_{q^-}, \ldots, z_{q^+})$ denote the joint density of

$$Z^-_{n,(q)} \le \cdots \le Z^-_{n,(1)} < 0 \le Z^+_{n,(1)} \le \cdots \le Z^+_{n,(q)} \ ,$$

so that we can write the last term in the previous display as

$$\int_0^\infty \int_0^{z_{q^+}} \cdots \int_0^{z_{(q-1)^-}} \Pi_{j=1}^q H(w_j^-|z_{j^-}) \cdot \Pi_{j=1}^q H(w_j^+|z_{j^+}) f_{n,(q^-,\ldots,q^+)}(z_{q^-},\ldots,z_{q^+}) dz_{q^-},\ldots,dz_{q^+} \ .$$

By (2.3), the integrand term

$$\Pi_{j=1}^q H(w_j^-|z_{j^-}) \cdot \Pi_{j=1}^q H(w_j^+|z_{j^+})$$

is a bounded continuous function of $(z_{q^-},\ldots,z_{1^-},z_{1^+},\ldots,z_{q^+})$ at $(0,0,\ldots,0)$. Suppose that the order statistics $Z_{n,(j)}^-$ and $Z_{n,(q)}^+$, for $j \in \{1,\ldots,q\}$, converge in distribution to a degenerate distribution with mass at $(0,0,\ldots,0)$. It would then follow from the definition of weak convergence, the asymptotic uniform integrability of the integrand term above, and van der Vaart (1998a, Theorem 2.20) that

$$\lim_{n\to\infty} E\left[\Pi_{j=1}^q H(w_j^-|z_{j^-}) \cdot \Pi_{j=1}^q H(w_j^+|z_{j^+})\right] = E\left[\Pi_{j=1}^q H^-(w_j^-|0) \cdot \Pi_{j=1}^q H^+(w_j^+|0)\right] \ .$$

Hence, it is sufficient to prove that for any given $j \in \{1,\ldots,q\}$, $Z_{n,(j)}^- = o_p(1)$ and $Z_{n,(q)}^+ = o_p(1)$. We prove $Z_{n,(q)}^+ = o_p(1)$ by complete induction, and omit the other proof as the result follows from similar arguments.

Take $j = 1$ and let $\epsilon > 0$. By Assumption 2.4.1, it follows that

$$F^+(\epsilon) = \Pr\{Z_i \in [0,\epsilon]\} > 0 \ .$$

Next, note that

$$\begin{aligned} F^{+}_{n,(1)}(\epsilon) &\equiv \Pr\{Z^{+}_{n,(1)} \leq \epsilon\} = \Pr\{\text{ at least 1 of the } Z_i \text{ is such that } Z_i \in [0,\epsilon]\} \\ &= \sum_{i=1}^{n} \binom{n}{i} [F^{+}(\epsilon)]^i [1-F^{+}(\epsilon)]^{n-i} \\ &= \sum_{i=0}^{n} \binom{n}{i} [F^{+}(\epsilon)]^i [1-F^{+}(\epsilon)]^{n-i} - [1-F^{+}(\epsilon)]^n \\ &= 1 - [1-F^{+}(\epsilon)]^n \ . \end{aligned} \tag{B.1}$$

Since $F^{+}(\epsilon) > 0$ for any $\epsilon > 0$, it follows that $\Pr\{Z^{+}_{n,(1)} > \epsilon\} = [1-F^{+}(\epsilon)]^n \to 0$ as $n \to \infty$ and $Z^{+}_{n,(1)} = o_p(1)$. Now let $F^{+}_{n,(j)}(\epsilon)$ denote the cdf of $Z^{+}_{n,(j)}$, which is given by

$$\begin{aligned} F^{+}_{n,(j)}(\epsilon) &= \Pr\{Z^{+}_{n,(j)} \leq \epsilon\} \\ &= \Pr\{\text{ at least } j \text{ of the } Z_i \text{ are such that } Z_i \in [0,\epsilon]\} \\ &= \sum_{i=j}^{n} \binom{n}{i} [F^{+}(\epsilon)]^i [1-F^{+}(\epsilon)]^{n-i} \\ &= F^{+}_{n,(j+1)}(\epsilon) + \binom{n}{j} [F^{+}(\epsilon)]^j [1-F^{+}(\epsilon)]^{n-j} \ , \end{aligned}$$

so that we can write

$$1 - F^{+}_{n,(j+1)}(\epsilon) = 1 - F_{n,(j)}(\epsilon) - \binom{n}{j} [F^{+}(\epsilon)]^j [1-F^{+}(\epsilon)]^{n-j} \text{ for } j \in \{1,\ldots,q-1\} \ . \tag{B.2}$$

It follows from (B.1) that $1 - F^{+}_{n,(1)}(\epsilon) \to 0$ for any $\epsilon > 0$ as $n \to \infty$. In order to complete the proof we assume that $1 - F^{+}_{n,(j)}(\epsilon) \to 0$ for $j \in \{1,\ldots,q-1\}$ and show that this

implies that $1 - F^{+}_{n,(j+1)}(\epsilon) \to 0$. By (B.2) this is equivalent to showing that

$$\binom{n}{j}[F^{+}(\epsilon)]^{j}[1 - F^{+}(\epsilon)]^{n-j} \to 0 \ .$$

To this end, note that

$$\binom{n}{j}[F^{+}(\epsilon)]^{j}[1 - F^{+}(\epsilon)]^{n-j} \leq n^{j}[1 - F^{+}(\epsilon)]^{n-j} = \left[e^{\frac{j \log n}{n-j}}[1 - F^{+}(\epsilon)]\right]^{n-j} \to 0 \ ,$$

where the convergence follows after noticing that there exists $N \in \mathbf{R}$ such that $e^{\frac{j \log n}{n-j}}[1 - F^{+}(\epsilon)] < 1$ for all $n > N$ and any $j \in \{1, \ldots, q-1\}$. The result follows. ■

B.0.2. Proof of Theorem 2.4.2

Part 1. Continuous case: Let $P_n = \otimes_{i=1}^{n} P$ with $P \in \mathbf{P}_0$ be given. By Assumption 2.4.4(i) and the Almost Sure Representation Theorem (see van der Vaart, 1998a, Theorem 2.19), there exists $\tilde{S}_n$, $\tilde{S}$, and $U \sim U(0,1)$, defined on a common probability space $(\Omega, \mathcal{A}, \tilde{P})$, such that

$$\tilde{S}_n \to \tilde{S} \text{ w.p.1} \ ,$$

$\tilde{S}_n \stackrel{d}{=} S_n$, $\tilde{S} \stackrel{d}{=} S$, and $U \perp (\tilde{S}_n, \tilde{S})$. Consider the permutation test based on $\tilde{S}_n$, this is,

$$\tilde{\phi}(\tilde{S}_n, U) \equiv \begin{cases} 1 & T(\tilde{S}_n) > T^{(k)}(\tilde{S}_n) \text{ or } T(\tilde{S}_n) = T^{(k)}(\tilde{S}_n) \text{ and } U < a(\tilde{S}_n) \\ 0 & T(\tilde{S}_n) < T^{(k)}(\tilde{S}_n) \end{cases} .$$

Denote the randomization test based on $\tilde{S}$ by $\tilde{\phi}(\tilde{S}, U)$, where the same uniform variable U is used in $\tilde{\phi}(\tilde{S}_n, U)$ and $\tilde{\phi}(\tilde{S}, U)$.

Since $\tilde{S}_n \stackrel{d}{=} S_n$, it follows immediately that $E_{P_n}[\phi(S_n)] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)]$. In addition, since $\tilde{S} \stackrel{d}{=} S$, Assumption 2.4.4(ii) implies that $E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] = \alpha$ by the usual arguments behind randomization tests, see Lehmann and Romano (2005, Chapter 15). It therefore suffices to show

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] \to E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] \ . \tag{B.3}$$

In order to show (B.3), let E_n be the event where the ordered values of $\{S_j : 1 \leq j \leq 2q\}$ and $\{S_{n,j} : 1 \leq j \leq 2q\}$ correspond to the same permutation π of $\{1, \ldots, 2q\}$, i.e., if $S_{\pi(j)} = S_k$ then $S_{n,\pi(j)} = S_{n,k}$ for $1 \leq j \leq 2q$ and $1 \leq k \leq 2q$. We first claim that $I\{E_n\} \to 1$ w.p.1. To see this, note that Assumption 2.4.4(iii) and $\tilde{S} \stackrel{d}{=} S$ imply that

$$\tilde{S}_{(1)}(\omega) < \tilde{S}_{(2)}(\omega) < \cdots < \tilde{S}_{(2q)}(\omega) \tag{B.4}$$

for all ω in a set with probability one under $\tilde{P}$. Moreover, since $\tilde{S}_n \to \tilde{S}$ w.p.1, there exists a set Ω^* with $\tilde{P}\{\Omega^*\} = 1$ such that both (B.4) and $\tilde{S}_n(\omega) \to \tilde{S}(\omega)$ hold for all $\omega \in \Omega^*$. For all ω in this set, let $\pi(1, \omega), \ldots, \pi(2q, \omega)$ be the permutation that delivers the order statistics in (B.4). It follows that for any $\omega \in \Omega^*$ and any $j \in \{1, \ldots, 2q-1\}$, if $\tilde{S}_{\pi(j,\omega)}(\omega) < \tilde{S}_{\pi(j+1,\omega)}(\omega)$ then

$$\tilde{S}_{n,\pi(j,\omega)}(\omega) < \tilde{S}_{n,\pi(j+1,\omega)}(\omega) \text{ for } n \text{ sufficiently large} \ . \tag{B.5}$$

We can therefore conclude that

$$I\{E_n\} \to 1 \ w.p.1 \ ,$$

which proves the first claim.

We now prove (B.3) in two steps. First, we note that

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] \ . \tag{B.6}$$

This is true because, on the event E_n, the rank statistics in (2.19) of the vectors $\tilde{S}_n^{\pi}$ and $\tilde{S}^{\pi}$ coincide for all $\pi \in \mathbf{G}$, and by Assumption 2.4.4(iv), the test statistic $T(S)$ only depends on the order of the observations, leading to $\tilde{\phi}(\tilde{S}_n, U) = \tilde{\phi}(\tilde{S}, U)$ on E_n. Second, since $I\{E_n\} \to 1$ w.p.1 it follows that $\tilde{\phi}(\tilde{S}, U)I\{E_n\} \to \tilde{\phi}(\tilde{S}, U)$ w.p.1 and $\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\} \to 0$ w.p.1. We can therefore use (B.6) and invoke the dominated convergence theorem to conclude that,

$$\begin{aligned} E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &\to E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] \ . \end{aligned}$$

This completes the proof of the first part of the statement of the theorem for the continuous case.

Discrete case: The proof for the discrete setting is similar to the continuous one with few intuitive differences. We reproduce it here for completeness.

Let $P_n = \otimes_{i=1}^n P$ with $P \in \mathbf{P}_0$ be given. By Assumption 2.4.5(i) and the Almost Sure Representation Theorem (see van der Vaart, 1998a, Theorem 2.19), there exists $\tilde{S}_n$, $\tilde{S}$,

and $U \sim U(0,1)$, defined on a common probability space $(\Omega, \mathcal{A}, \tilde{P})$, such that

$$\tilde{S}_n \to \tilde{S} \text{ w.p.1 } ,$$

$\tilde{S}_n \stackrel{d}{=} S_n,\ \tilde{S} \stackrel{d}{=} S,$ and $U \perp (\tilde{S}_n, \tilde{S})$. Consider the permutation test based on $\tilde{S}_n$, this is,

$$\tilde{\phi}(\tilde{S}_n, U) \equiv \begin{cases} 1 & T(\tilde{S}_n) > T^{(k)}(\tilde{S}_n) \text{ or } T(\tilde{S}_n) = T^{(k)}(\tilde{S}_n) \text{ and } U < a(\tilde{S}_n) \\ 0 & T(\tilde{S}_n) < T^{(k)}(\tilde{S}_n) \end{cases} .$$

Denote the randomization test based on $\tilde{S}$ by $\tilde{\phi}(\tilde{S}, U)$, where the same uniform variable U is used in $\tilde{\phi}(\tilde{S}_n, U)$ and $\tilde{\phi}(\tilde{S}, U)$.

Since $\tilde{S}_n \stackrel{d}{=} S_n$, it follows immediately that $E_{P_n}[\phi(S_n)] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)]$. In addition, since $\tilde{S} \stackrel{d}{=} S$, Assumption 2.4.5(ii) implies that $E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] = \alpha$ by the usual arguments behind randomization tests, see Lehmann and Romano (2005, Chapter 15). It therefore suffices to show

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] \to E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] \ . \tag{B.7}$$

In order to show (B.7), let E_n be the event where $\tilde{S}_n = \tilde{S}$. We first claim that $I\{E_n\} \to 1$ w.p.1. To see this, note that by Assumption 2.4.5(iii), the discrete random variable $\tilde{S}_n$ takes values in $\mathcal{S}_n \subseteq \mathcal{S} \equiv \otimes_{j=1}^{2q} \mathcal{S}_1$ for all $n \geq 1$. The set $\mathcal{S}$ is closed by virtue of being a finite collection of singletons, and by the Portmanteau Lemma (see van der Vaart, 1998a, Lemma 2.2) it follows that

$$1 = \limsup_{n\to\infty} \tilde{P}\{\tilde{S}_n \in \mathcal{S}_n\} \leq \limsup_{n\to\infty} \tilde{P}\{\tilde{S}_n \in \mathcal{S}\} \leq \tilde{P}\{\tilde{S} \in \mathcal{S}\} \ , \tag{B.8}$$

meaning that $\text{supp}(\tilde{S}) \subseteq \mathcal{S}$. Moreover, since $\tilde{S}_n \to \tilde{S}$ w.p.1, there exists a set Ω^* with $\tilde{P}\{\Omega^*\} = 1$ such that $\tilde{S}_n(\omega) \to \tilde{S}(\omega)$ holds for all $\omega \in \Omega^*$. It follows that for any $\omega \in \Omega^*$ and any $j \in \{1, \ldots, 2q\}$,

$$\tilde{S}_{n,j}(\omega) = \tilde{S}_j(\omega) \text{ for } n \text{ sufficiently large} \quad , \tag{B.9}$$

which follows from the fact that both $\tilde{S}$ and $\tilde{S}_n$ are discrete random variables taking values in (possibly a subset of) the finite collection of points in $\mathcal{S} \equiv \otimes_{j=1}^{2q} \mathcal{S}_1 = \otimes_{j=1}^{2q} \bigcup_{k=1}^{m} \{a_k\}$. We conclude that

$$I\{E_n\} \to 1 \ w.p.1 \ ,$$

which proves the first claim.

We now prove (B.7) in two steps. First, we note that

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U) I\{E_n\}] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U) I\{E_n\}] \ . \tag{B.10}$$

This is true because, on the event E_n, $\tilde{S}_n^\pi$ and $\tilde{S}^\pi$ coincide for all $\pi \in \mathbf{G}$, leading to $\tilde{\phi}(\tilde{S}_n, U) = \tilde{\phi}(\tilde{S}, U)$ on E_n. Second, since $I\{E_n\} \to 1$ w.p.1 it follows that $\tilde{\phi}(\tilde{S}, U) I\{E_n\} \to \tilde{\phi}(\tilde{S}, U)$ w.p.1 and $\tilde{\phi}(\tilde{S}_n, U) I\{E_n^c\} \to 0$ w.p.1. We can therefore use (B.10) and invoke the dominated convergence theorem to conclude that,

$$\begin{aligned} E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U) I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U) I\{E_n^c\}] \\ &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U) I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U) I\{E_n^c\}] \\ &\to E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] \ . \end{aligned}$$

This completes the proof for the discrete case and the first part of the statement of the theorem.

Part 2. Let $P_n = \otimes_{i=1}^{n} P$ with $P \in \mathbf{P}_0$ be given and note that by Theorem 2.4.1 it follows that

$$\begin{aligned} S_n &= (S_{n,1}, \ldots, S_{n,2q}) = (W^-_{n,[1]}, \ldots, W^-_{n,[q]}, W^+_{n,[1]}, \ldots, W^+_{n,[q]}) \\ &\xrightarrow{d} (S_1, \ldots, S_{2q}) \ , \end{aligned}$$

where $(S_1, \ldots, S_{2q})$ are i.i.d. with cdf $H(w|0)$. The conditions in Assumption 2.4.4.(i)-(ii) immediately follow as $(S_1, \ldots, S_{2q}) \stackrel{d}{=} (S_{\pi(1)}, \ldots, S_{\pi(2q)})$ for any $\pi \in \mathbf{G}$. Assumption 2.4.4.(iii) follows the fact that $(S_1, \ldots, S_{2q})$ are i.i.d. with cdf $H(w|0)$, where $H(w|0)$ is the cdf of a continuous random variable by Assumption 2.4.2. Similarly, Assumption 2.4.5.(iii) follows the fact that $(S_1, \ldots, S_{2q})$ are i.i.d. with cdf $H(w|0)$, where $H(w|0)$ is the cdf of a discrete random variable by Assumption 2.4.3.

We are left to prove that the test statistic in (2.12) satisfies Assumption 2.4.4.(iv). To show this, note that $T(S)$ as in (2.12) admits the alternative representation

$$T(S) = \frac{1}{q} T^*(S) - \frac{4q^2 - 1}{12q} \ ,$$

where

$$T^*(S) = \frac{1}{q} \sum_{i=1}^{q} (R^*_i - i)^2 + \frac{1}{q} \sum_{j=1}^{q} (R^*_{q+j} - j)^2 \ ,$$

$R^*_1 < R^*_2 < \cdots < R^*_q$ denote the increasingly ordered ranks $R_1, \ldots, R_q$ of the first q variables in S, and $R^*_{q+1} < \cdots < R^*_{2q}$ are the increasingly ordered ranks $R_{q+1}, \ldots, R_{2q}$ of

the last q values in S. It follows immediately that this test statistic satisfies Assumption 2.4.4.(iv). This completes the proof of the second part of the statement of the theorem. ■

B.0.3. The multidimensional case

In this appendix we discuss the case where W is a K-dimensional vector. The test statistic in (2.12) and the test construction in (2.14) immediately apply to this case where W is a vector consisting of a combination of discrete and continuously distributed random variables. However, in the multidimensional case we also consider an alternative test statistic that may exhibit better power when W includes several components and some are continuous and some are discontinuous at the threshold. We call this test statistic the max test statistic and define it as follows,

$$T_{\max}(S_n) = \max_{c \in \hat{\mathbf{C}}} T(c'S_n) \ , \tag{B.11}$$

where $T(\cdot)$ is the test statistic in (2.12),

$$c'S_n = (c'S_{n,1}, \ldots, c'S_{n,2q}) = (c'W^-_{n,[1]}, \ldots, c'W^-_{n,[q]}, c'W^+_{n,[1]}, \ldots, c'W^+_{n,[q]}) \ , \tag{B.12}$$

and $\hat{\mathbf{C}}$ is a collection of elements from the unit sphere $\mathbf{C} \equiv \{c \in \mathbf{R}^K : ||c|| = 1\}$. The intuition behind this test statistic arises from observing that the null hypothesis in (2.3) is equivalent to the same statement applied to any univariate projection $c'W$ of W, i.e.,

$$\Pr\{c'W \leq w | Z = z\} \text{ is continuous in } z \text{ at } z = 0 \text{ for all } w \in \mathbf{R} \text{ and all } c \in \mathbf{C}. \tag{B.13}$$

In the empirical application of Section 2.6 we choose $\hat{\mathbf{C}}$ to include $100 - K$ i.i.d. draws from Uniform($\mathbf{C}$) together with the K canonical elements (i.e., vectors c with zeros in all coordinates except for one). We also set $\hat{q}_{\text{rot}}$ to be the minimum value across the rule of thumb across each individual covariate, i.e., $\hat{q}_{\text{rot}} = \min\{\hat{q}_{\text{rot,1}}, \ldots, \hat{q}_{\text{rot,K}}\}$.

Given a test statistic, here we show that the permutation test for this setting is also asymptotically valid. We first state the primitive conditions required to prove this.

Assumption B.0.1. *For any $\epsilon > 0$, Z satisfies $\Pr\{Z \in (-\epsilon, 0)\} > 0$ and $\Pr\{Z \in [0, \epsilon]\} > 0$.*

Assumption B.0.2. *The random vector W takes values in $\mathbf{R}^{d_w}$ and has components W_k, for $k \in \{1, \ldots, d_w\}$, that satisfy either Assumption 2.4.2 or Assumption 2.4.3 with $|\mathcal{W}_k| = m_k \in \mathbf{N}$ points of support.*

Assumption B.0.1 is the same as Assumption 2.4.1, which is required for Theorem 2.4.1 to hold. Moreover, Assumption B.0.2 essentially requires that each component of the vector W satisfies one of the two assumptions we used for the scalar case.

We formalize the high level assumptions required for the validity of the permutation test for the vector case in the following assumption.

Assumption B.0.3. *If $P \in \mathbf{P}_0$, then*

(i) $S_n = S_n(X^{(n)}) \xrightarrow{d} S$ *under* P_n.

(ii) $S^\pi \overset{d}{=} S$ *for all* $\pi \in \mathbf{G}$.

(iii) $S_n = (S_{n,1}, \ldots, S_{n,2q})$ *is such that each* $S_{n,j}$, $j \in \{1, \ldots, 2q\}$, *takes values in* $\mathbf{R}^{d_w}$ *and has single components* $S_{n,j,k}$, $k \in \{1, \ldots, d_w\}$, *that are either continuously distributed taking values in* $\mathbf{R}$ *or discretely distributed taking values* $\mathcal{S}_{n,k} \subseteq S_1 = \bigcup_{\ell=1}^{m} \{a_\ell\}$

with $a_\ell \in \mathbf{R}$ *distinct. In addition, for each component* $S_{n,j,k}$, $k \in \{1, \ldots, d_w\}$, *that is continuously distributed, the corresponding component in* $S = (S_1, \ldots, S_{2q})$, $S_{j,k}$, *is also continuously distributed.*

(iv) $T : \mathcal{S} \to \mathbf{R}$ *is invariant to rank with respect to each continuous component, i.e., it only depends on the order of the elements of each continuous component.*

We now formalize our result for the vector case in Theorem B.0.1, which shows that the permutation test defined in (2.14) leads to a test that is asymptotically level α whenever Assumption B.0.3 holds. In addition, the same theorem also shows that Assumption B.0.1-B.0.2 are sufficient primitive conditions for the asymptotic validity of our test.

Theorem B.0.1. *Suppose that Assumption B.0.3 holds and let* $\alpha \in (0, 1)$. *Then,* $\phi(S_n)$ *defined in* (2.14) *satisfies*

$$E_P[\phi(S_n)] \to \alpha \tag{B.14}$$

as $n \to \infty$ *whenever* $P \in \mathbf{P}_0$. *Moreover, if* $T : \mathcal{S} \to \mathbf{R}$ *is the Cramér Von Mises test statistic in* (2.12) *and Assumptions B.0.1-B.0.2 hold, then Assumption B.0.3 also holds and* (B.14) *follows.*

B.0.3.1. Proof of Theorem B.0.1.

Part 1. Let $P_n = \otimes_{i=1}^n P$ with $P \in \mathbf{P}_0$ be given. By Assumption B.0.3(i) and the Almost Sure Representation Theorem (see van der Vaart, 1998a, Theorem 2.19), there exists $\tilde{S}_n$, $\tilde{S}$, and $U \sim U(0, 1)$, defined on a common probability space $(\Omega, \mathcal{A}, \tilde{P})$, such that

$$\tilde{S}_n \to \tilde{S} \text{ w.p.1 },$$

$\tilde{S}_n \stackrel{d}{=} S_n,\ \tilde{S} \stackrel{d}{=} S,$ and $U \perp (\tilde{S}_n, \tilde{S})$. Consider the permutation test based on $\tilde{S}_n$, this is,

$$\tilde{\phi}(\tilde{S}_n, U) \equiv \begin{cases} 1 & T(\tilde{S}_n) > T^{(k)}(\tilde{S}_n) \text{ or } T(\tilde{S}_n) = T^{(k)}(\tilde{S}_n) \text{ and } U < a(\tilde{S}_n) \\ 0 & T(\tilde{S}_n) < T^{(k)}(\tilde{S}_n) \end{cases} .$$

Denote the randomization test based on $\tilde{S}$ by $\tilde{\phi}(\tilde{S}, U)$, where the same uniform variable U is used in $\tilde{\phi}(\tilde{S}_n, U)$ and $\tilde{\phi}(\tilde{S}, U)$.

Since $\tilde{S}_n \stackrel{d}{=} S_n$, it follows immediately that $E_{P_n}[\phi(S_n)] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)]$. In addition, since $\tilde{S} \stackrel{d}{=} S$, Assumption B.0.3(ii) implies that $E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] = \alpha$ by the usual arguments behind randomization tests, see Lehmann and Romano (2005, Chapter 15). It therefore suffices to show

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] \to E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] \ . \tag{B.15}$$

Before we show (B.15), we introduce the additional notation to easily refer to the different components of the vectors S_j and $S_{n,j}$ for $j \in \{1, \ldots, 2q\}$. Let the first K^c elements of S_j and $S_{n,j}$ denote the continuous components, where each component is denoted by $S^c_{j,k}$ and $S^c_{n,j,k}$ for $1 \leq k \leq K^c$. Let the remaining subvector S^d_j and $S^d_{n,j}$ of dimension $K^d = K - K^c$ denote the discrete component of S_j and $S_{n,j}$. Arguing as in the proof of Theorem 2.4.2, it follows that S^d_j has support in (a possible subset of) the same finite collection of points that $S^d_{n,j}$ may take values. For simplicity here and wlog, denote by $(s^*_1, \ldots, s^*_L)$ the common points of support of S^d_j and $S^d_{n,j}$. Using this notation, we can partition S_j and $S_{n,j}$ as (S^c_j, S^d_j) and $(S^c_{n,j}, S^d_{n,j})$, respectively.

In order to show (B.15), let E_n be the event where the following holds. First, the ordered values of each continuous component $\{S^c_{j,k} : 1 \leq j \leq 2q\}$ and $\{S^c_{n,j,k} : 1 \leq j \leq 2q\}$ correspond to the same permutation π_k of $\{1, \ldots, 2q\}$ for $1 \leq k \leq K^c$, i.e., if $S^c_{k,\pi_k(j)} = S^c_{k,l}$ then $S^c_{n,k,\pi_k(j)} = S^c_{n,k,l}$ for $1 \leq j, l \leq 2q$ and $1 \leq k \leq K^c$. Second, the discrete subvectors $\{S^d_j : 1 \leq j \leq 2q\}$ and $\{S^d_{n,j} : 1 \leq j \leq 2q\}$ coincide, i.e., $S^d_j = S^d_{n,j}$ for $1 \leq j \leq 2q$.

We first claim that $I\{E_n\} \to 1$ w.p.1. To see this, note that Assumption B.0.3(iii) and $\tilde{S} \stackrel{d}{=} S$ imply that for all ω in a set with probability one under $\tilde{P}$ we have for each continuous component k of S that

$$\tilde{S}^c_{k,(1)}(\omega) < \tilde{S}^c_{k,(2)}(\omega) < \cdots < \tilde{S}^c_{k,(2q)}(\omega) \ , \tag{B.16}$$

and for the discrete subvector of $\tilde{S}$ that

$$\tilde{S}^d_j(\omega) = s^*_l \ , \tag{B.17}$$

for $1 \leq j \leq 2q$ and some $1 \leq l \leq L$. Moreover, since $\tilde{S}_n \to \tilde{S}$ w.p.1, there exists a set Ω^* with $\tilde{P}\{\Omega^*\} = 1$ such that (B.16), (B.17) and $\tilde{S}_n(\omega) \to \tilde{S}(\omega)$ hold for all $\omega \in \Omega^*$. For all ω in this set, let $\pi_k(1, \omega), \ldots, \pi_k(2q, \omega)$ be the permutation that delivers the order statistics in (B.16) for the k^{th} continuous component. It follows that for any $\omega \in \Omega^*$ and any $j \in \{1, \ldots, 2q-1\}$, if for any continuous component k we have $\tilde{S}^c_{k,\pi_k(j,\omega)}(\omega) < \tilde{S}^c_{k,\pi_k(j+1,\omega)}(\omega)$ then

$$\tilde{S}^c_{n,k,\pi_k(j,\omega)}(\omega) < \tilde{S}^c_{n,k,\pi_k(j+1,\omega)}(\omega) \text{ for } n \text{ sufficiently large} \ , \tag{B.18}$$

and moreover, if for the discrete subvector we have $\tilde{S}_j^d(\omega) = s_l^*$ then

$$\tilde{S}_{n,j}^d(\omega) = s_l^* \text{ for } n \text{ sufficiently large} \quad , \tag{B.19}$$

which follows from the fact that both $\{S_j^d : 1 \leq j \leq 2q\}$ and $\{S_{n,j}^d : 1 \leq j \leq 2q\}$ are discretely distributed with common support points. We can therefore conclude that

$$I\{E_n\} \to 1 \ w.p.1 \ ,$$

which proves the first claim.

We now prove (B.15) in two steps. First, we note that

$$E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U) I\{E_n\}] = E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U) I\{E_n\}] \ . \tag{B.20}$$

This is true because, on the event E_n, the following two hold. First, for each continuous component the rank statistics in (2.19) of the vectors $\tilde{S}_{n,k}^{c,\pi}$ and $\tilde{S}_k^{c,\pi}$ coincide for $1 \leq k \leq K^c$ and for all $\pi \in \mathbf{G}$. Then we have by Assumption B.0.3(iv) that the test statistic $T(S)$ only depends on the order of the elements of each continuous component. Second, the discrete subvectors $\tilde{S}_n^{d,\pi}$ and $\tilde{S}^{d,\pi}$ coincide for all $\pi \in \mathbf{G}$. These two properties in turn result in, on the event E_n, $T(\tilde{S}_n^\pi)$ equaling $T(\tilde{S}^\pi)$ for all $\pi \in \mathbf{G}$, which leads to $\tilde{\phi}(\tilde{S}_n, U) = \tilde{\phi}(\tilde{S}, U)$ on E_n.

Then for the second step in proving (B.15), since $I\{E_n\} \to 1$ w.p.1 it follows that $\tilde{\phi}(\tilde{S}, U) I\{E_n\} \to \tilde{\phi}(\tilde{S}, U)$ w.p.1 and $\tilde{\phi}(\tilde{S}_n, U) I\{E_n^c\} \to 0$ w.p.1. We can therefore use

(B.20) and invoke the dominated convergence theorem to conclude that,

$$\begin{aligned} E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)] &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &= E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)I\{E_n\}] + E_{\tilde{P}}[\tilde{\phi}(\tilde{S}_n, U)I\{E_n^c\}] \\ &\to E_{\tilde{P}}[\tilde{\phi}(\tilde{S}, U)] \ . \end{aligned}$$

This completes the proof of the first part of the statement of the theorem.

Part 2. Let $P_n = \otimes_{i=1}^n P$ with $P \in \mathbf{P}_0$ be given and note that by Theorem 2.4.1 it follows that

$$\begin{aligned} S_n = (S_{n,1}, \ldots, S_{n,2q}) &= (W^-_{n,[1]}, \ldots, W^-_{n,[q]}, W^+_{n,[1]}, \ldots, W^+_{n,[q]}) \\ &\xrightarrow{d} (S_1, \ldots, S_{2q}) \ , \end{aligned}$$

where $(S_1, \ldots, S_{2q})$ are i.i.d. with cdf $H(w|0)$. The conditions in Assumption B.0.3.(i)-(ii) immediately follow as $(S_1, \ldots, S_{2q}) \stackrel{d}{=} (S_{\pi(1)}, \ldots, S_{\pi(2q)})$ for any $\pi \in \mathbf{G}$. Assumption B.0.3.(iii) also follows immediately by Assumption B.0.2. Finally, to show Assumption B.0.3.(iv) we first demonstrate that the test statistic in (2.12) admits an alternate representation. By Assumption B.0.2, let without loss of generality the first K^c components be continuous and the rest be discrete. Denote by S_i^d the discrete subvector of S_i and by

$$R_{i,k} = \sum_{j=1}^{2q} I\{S_{j,k}^c \leq S_{i,k}^c\} \ ,$$

the rank of the k^{th} continuous component of S_i for $1 \leq i \leq 2q$ and $1 \leq k \leq K^c$. Finally, the test statistic can be rewritten in the following alternate representation

$$T(S) = \frac{1}{2q} \sum_{j=1}^{2q} \left(\frac{1}{q} \sum_{i=1}^{q} \left[I\{S_i^d \leq S_j^d\} \prod_{k=1}^{K^c} 1\{R_{i,k} \leq R_{j,k}\} \right] - \frac{1}{q} \sum_{i=q+1}^{2q} \left[I\{S_i^d \leq S_j^d\} \prod_{k=1}^{K^c} \{R_{i,k} \leq R_{j,k}\} \right] \right)^2 .$$

The above representation follows from first rewriting

$$I\{S_i \leq S_j\} = I\{S_i^d \leq S_j^d\} \prod_{k=1}^{K^c} I\{S_{i,k}^c \leq S_{j,k}^c\} ,$$

and then noticing that for $1 \leq k \leq K^c$

$$I\{S_{i,k}^c \leq S_{j,k}^c\} = I\{R_{i,k} \leq R_{j,k}\} .$$

This representation illustrates that for the continuous components the test statistic only depends on their individual orderings. It then follows immediately that this test statistic satisfies Assumption B.0.3.(iv). This completes the proof of the second part of the statement of the theorem. ■

B.0.4. Additional details on the simulations

In this appendix we document some computational details on the simulations of Section 2.5. The Matlab codes to replicate all our results are available online and include a discussion on the details mentioned here.

Details on $\hat{q}_{\text{rot}}$. The feasible rule of thumb for q is computed (in our simulations and in the companion Stata package) as follows:

$$\hat{q}_{\text{rot}} = \left\lceil \max \left\{ \min \left\{ \hat{f}_n(0) \hat{\sigma}_{Z,n} \left(1 - \hat{\rho}_n^2\right)^{1/2} \frac{n^{0.9}}{\log n} \ , \ q_{UB} \right\} \ , \ q_{LB} \right\} \right\rceil \ ,$$

where q_{LB} and q_{UB} are a lower and upper bounds, respectively. We set $q_{LB} = 10$, as less than 10 observations leads to tests where the randomized and non-randomized versions of the permutation test differ. We then set $q_{UB} = \frac{n^{0.9}}{\log n}$, as $\frac{n}{\log n}$ is the rate that violates the conditions we require for q in the proof of Theorem 2.4.1. The estimator $\hat{f}_n(0)$ of $f(0)$ is a kernel estimator with a triangular kernel and a bandwidth h computed using Silverman's rule of thumb. The estimators $\hat{\rho}_n$ and $\hat{\sigma}^2_{Z,n}$ are the sample correlation between W_i and Z_i and sample variance of Z_i.

Details on SZ bandwidth. Shen and Zhang (2016) propose the rule of thumb bandwidth in (2.23), where h_n^{CCT} is a two step bandwidth estimate based on Calonico et al. (2014b). In the first step, a pilot bandwidth is selected using CCT for estimating the average treatment effect at the cutoff. Note that this is the same bandwidth used in the CCT test. Then, in the second step, CCT is used again with the dependent variables as $I\{W_i \leq \tilde{w}\}$, where $\tilde{w}$ corresponds to the minimum amongst the values that attain the maximum estimated distributional treatment effect. In our simulations, however, this results in no variation in the dependent variable in some models, which leads to the termination of the program. In such cases when there is no variation, for example Model 6, we first take $\tilde{w}$ to be the estimated median value of W_i using the whole sample of data. If this additionally fails, we take h_n^{CCT} to be the pilot bandwidth. Shen and Zhang (2016) additionally propose an alternative rule of thumb based on the bandwidth proposed by Imbens and Kalyanaraman (2012a)., and find similar results. We hence do not include results of this alternative choice in our comparisons.

We finally note that in the cases were W is discrete (either in our simulations or in the empirical application) we implemented test SZ as described in Section 2.5. Shen and

Zhang (2016) mention an alternative implementation of their test based on the Wald test statistic and a bootstrap critical value, but do not provide further details on the implementation of such a variation. For this reason, we use the same test regardless of whether W is continuous or discrete.

B.0.5. Surveyed papers on RDD

Table B.1 displays the list of papers we surveyed in leading journals that use regression discontinuity designs. We specifically note whether these papers test for any of the two implications we mention in the introduction, namely, validating the continuity of the density of the running variable and validating the continuity of the means of the baseline covariates.

We briefly describe the criteria used to prepare our list. The journals selected were the *American Economic Review* (AER), the *American Economic Journal: Applied Economics* (AEJ:AppEcon), the *Quarterly Journal of Economics* (QJE), and the *Review of Economics and Statistics* (ReStat), and the years used were from the beginning of 2011 to the end of 2015. All papers in each volumes were surveyed with the exception of the May volume for AER. We first categorized papers using regression discontinuity methods by searching the main text for the keywords 'regression discontinuity'. We then individually inspected the papers along with their appendices for whether they validated their design, and, if so, by either checking the continuity of the density of the running variable or the continuity of the means of the baseline covariates, or both. We allowed for both formal test results as well as informal graphical evidence.

Authors (Year)	Journal	Density Test	Mean Test	Authors (Year)	Journal	Density Test	Mean Test
Schmieder et al. (2016)	AER	✓	✓	Miller et al. (2013)	AEJ:AppEcon	✓	✓
Feldman et al. (2016)	AER	✓	✓	Litschig and Morrison (2013)	AEJ:AppEcon	✓	✓
Jayaraman et al. (2016)	AER	×	×	Dobbie and Skiba (2013)	AEJ:AppEcon	✓	✓
Dell (2015)	AER	✓	✓	Kazianga et al. (2013)	AEJ:AppEcon	✓	✓
Hansen (2015)	AER	✓	✓	Magruder (2012)	AEJ:AppEcon	×	×
Anderson (2014)	AER	×	×	Dustmann and Schnberg (2012)	AEJ:AppEcon	×	×
Martin et al. (2014)	AER	×	×	Clots-Figueras (2012)	AEJ:AppEcon	✓	✓
Dahl et al. (2014)	AER	✓	✓	Manacorda et al. (2011)	AEJ:AppEcon	✓	✓
Shigeoka (2014)	AER	×	✓	Chetty et al. (2014)	QJE	✓	✓
Crost et al. (2014)	AER	×	✓	Michalopoulos and Papaioannou (2014)	QJE	×	✓
Kostol and Mogstad. (2014)	AER	✓	✓	Fredriksson et al. (2013)	QJE	✓	✓
Clark and Royer (2013)	AER	×	✓	Schmieder et al. (2012)	QJE	✓	✓
Brollo et al. (2013)	AER	✓	✓	Lee and Mas (2012)	QJE	×	×
Bharadwaj et al. (2013)	AER	✓	✓	Saez et al. (2012)	QJE	×	×
Pop-Eleches and Urquiola (2013)	AER	✓	✓	Barreca et al. (2011)	QJE	×	×
Lacetera et al. (2012)	AER	✓	×	Almond et al. (2011)	QJE	✓	✓
Duflo et al. (2012)	AER	×	×	Malamud and Pop-Eleches (2011)	QJE	✓	✓
Gopinath et al. (2011)	AER	✓	✓	Fulford (2015)	ReStat	×	✓
Auffhammer and Kellogg (2011)	AER	×	×	Snider and Williams (2015)	ReStat	×	×
Duflo et al. (2011)	AER	×	×	Doleac and Sanders (2015)	ReStat	×	×
Ferraz and Finan (2011)	AER	×	×	Coşar et al. (2015)	ReStat	×	×
McCrary and Royer (2011)	AER	×	✓	Avery and Brevoort (2015)	ReStat	×	×
Beland (2015)	AEJ:AppEcon	✓	✓	Carpenter and Dobkin (2015)	ReStat	×	✓
Buser (2015)	AEJ:AppEcon	✓	✓	Black et al. (2014)	ReStat	✓	✓
Fack and Grenet (2015)	AEJ:AppEcon	✓	✓	Anderson et al. (2014)	ReStat	×	×
Cohodes and Goodman (2014)	AEJ:AppEcon	✓	✓	Alix-Garcia et al. (2013)	ReStat	×	✓
Haggag and Paci (2014)	AEJ:AppEcon	✓	✓	Albouy (2013)	ReStat	×	×
Dobbie and Fryer (2014)	AEJ:AppEcon	✓	✓	Garibaldi et al. (2012)	ReStat	✓	✓
Sekhri (2014)	AEJ:AppEcon	✓	✓	Manacorda (2012)	ReStat	✓	✓
Schumann (2014)	AEJ:AppEcon	✓	✓	Martorell and McFarlin (2011)	ReStat	✓	✓
Lucas and Mbiti (2014)	AEJ:AppEcon	✓	✓	Grosjean and Senik (2011)	ReStat	×	×

Table B.1. Papers using manipulation/placebo tests from 2011 − 2015.

We find that out of the 62 papers that use regression discontinuity methods, 35 validate by checking the continuity of the density, 42 validate by checking continuity of the baseline covariates, 34 validate using both tests, and 19 do not include any form of manipulation or placebo test.

APPENDIX C

Appendix to Chapter 3

C.0.1. Additional Notation

Let $Z^{(n)} = \{Z_i : 1 \leq i \leq n\}$ denote the observed sample of the random variable Z. Let $a_n \precsim b_n$ denote $a_n \leq Ab_n$, where a_n and b_n are deterministic sequences and A is a positive constant uniform in $\mathbf{P}$. Let $|\cdot|$ denote the Euclidean matrix norm. As we use the notion of convergence in probability under the sequence of distributions P_n, let $A_n = o_{P_n}(1)$ denote

$$P_n(|A_n| > \epsilon) \to 0 \text{ as } n \to \infty \ ,$$

for a sequences of random variables $A_n \sim P_n$, where ϵ is any constant such that $\epsilon > 0$. Further, in Table C.1 below, we introduce additional notation to keep our arguments concise.

$H(h_n)$	$\text{diag}(1,h_n^{-1},h_n^{-2})$
$r(Z_i/h_n)$	$(1, Z_i/h_n, (Z_i/h_n)^2)'$
$Z_n(h_n)$	$(r(Z_1/h_n), \ldots, r(Z_n/h_n))'$
$k(u)$	$(1-u)1\{0 \leq u \leq 1\}$
$K(u)$	$k(-u)1\{u < 0\} + k(u)1\{u \geq 0\}$
$K_{h_n}(u)$	$K(u/h_n)/h_n$
$W_{+,n}(h_n)$	$\text{diag}(1\{Z_1 \geq 0\}K_{h_n}(Z_1), \ldots, 1\{Z_n \geq 0\}K_{h_n}(Z_n))$

$\Gamma_{+,n}(h_n)$	$Z_n(h_n)'W_{+,n}(h_n)Z_n(h_n)/n$
$S_n(h_n)$	$((Z_1/h_n)^3, \ldots, (Z_n/h_n)^3)'$
$\nu_{+,n}$	$Z_n(h_n)'W_{+,n}(h_n)S_n(h_n)/n$
e	$(1,0,0)'$
$\mu(z,P)$	$E_P[Y\|Z=z]$
$\mu_+(P)$	$\lim_{z\to 0^+}\mu(z,P)$
$\mu_-(P)$	$\lim_{z\to 0^-}\mu(z,P)$
$\mu^v(z,P)$	$d^v\mu(z,P)/dz^v$
$\mu_+^v(P)$	$\lim_{z\to 0^+}\mu^v(z,P)$
$\sigma^2(z,P)$	$Var_P[Y\|Z=z]$
$\Sigma_n(P)$	$\mathrm{diag}(\sigma^2(Z_1,P),\ldots,\sigma^2(Z_n,P))$
$\Psi_{+,n}(h_n,P)$	$Z_n(h_n)'W_{+,n}(h_n)\Sigma_n(P)W_{+,n}(h_n)Z_n(h_n)/n$
$\mathbf{Y}_n$	$(Y_1,\ldots,Y_n)'$
$\hat{\beta}_{+,n}$	$H(h_n)\Gamma_{+,n}^{-1}(h_n)Z_n(h_n)'W_{+,n}(h_n)\mathbf{Y}_n/n$

Table C.1. Important Notation

Next, we provide an extended description of the test statistic used. For our null hypothesis as stated in the paper, the test statistic can be rewritten as

$$T_n^{CCT} = \frac{\hat{\mu}_{+,n} + \hat{\mu}_{-,n} - (\mu_+(P) - \mu_-(P))}{\hat{S}_n}\,, \tag{C.1}$$

where $\mu_+(P) - \mu_-(P) = \theta_0$, $\hat{\mu}_{+,n}$ and $\hat{\mu}_{-,n}$ are bias corrected local linear estimates of $\mu_+(P)$ and $\mu_-(P)$, and

$$\hat{S}_n = \sqrt{\hat{V}_{+,n} + \hat{V}_{-,n}} \ ,$$

where $\hat{V}_{+,n}$ and $\hat{V}_{+,n}$ are plug-in estimates conditional on $Z^{(n)}$ of the variances of $\hat{\mu}_{+,n}$ and $\hat{\mu}_{-,n}$; see (C.13) for the plug-in estimator used. The bias of both estimates are estimated using local quadratic estimators. Furthermore, for all estimates, we use the triangular kernel, i.e. $k(u)$ in Table C.1, and a deterministic sequence of bandwidth choices denoted by h_n. Throughout this document, we provide results for quantities with subscript $(+)$ as arguments for those with subscript $(-)$ follow symmetrically. In addition, as noted in Calonico et al. (2014a, Remark 7), we exploit the fact that in our simple version of the test statistic the estimates are numerically equivalent to those from a non-bias-corrected local quadratic estimator. In turn, we can write

$$\hat{\mu}_{+,n} = e'\hat{\beta}_{+,n} \ , \tag{C.2}$$

which reduces the length of the proof presented below. Further, as stated in the paper, note that

$$\mathbf{Q} = \{Q \in \mathbf{Q}_{\mathcal{W}} : Q \text{ satisfies Assumption 4.1}\} \ , \tag{C.3}$$

and that

$$\mathbf{P} = \{QM^{-1} : Q \in \mathbf{Q}\} \ , \tag{C.4}$$

where $\mathbf{Q}_{\mathcal{W}}$, M^{-1} and Assumption 4.1 are as defined in the paper.

C.0.2. Auxiliary Lemmas

Lemma C.0.1. *Let* $\mathbf{Q}$ *be defined as in* (C.3), $\mathbf{P}$ *be as in* (C.4) *and* $P_n \in \mathbf{P}$ *for all* $n \geq 1$. *If* $nh_n \to \infty$ *and* $h_n \to 0$, *then*

(i) $\Gamma_{+,n}(h_n) = \tilde{\Gamma}_{+,n}(h_n) + o_{P_n}(1)$, *where* $\tilde{\Gamma}_{+,n}(h_n) = \int_0^\infty K(u)r(u)r(u)'f_{P_n}(uh_n)du \in [\Gamma_L, \Gamma_U]$.

(ii) $\nu_{+,n}(h_n) = \tilde{\nu}_{+,n}(h_n) + o_{P_n}(1)$, *where* $\tilde{\nu}_{+,n}(h_n) = \int_0^\infty K(u)r(u)u^2 f_{P_n}(uh_n)du \in [\nu_L, \nu_U]$.

(iii) $h_n\Psi_{+,n}(h_n, P_n) = \tilde{\Psi}_{+,n}(h_n) + o_{P_n}(1)$, *where*

$$\tilde{\Psi}_{+,n}(h_n) = \int_0^\infty K(u)^2 r(u)r(u)'\sigma^2_{P_n}(uh_n)f_{P_n}(uh_n)du \in [\Psi_L, \Psi_U] \ .$$

Proof. For (i), a change of variables gives us

$$\begin{aligned} E_{P_n^n}[\Gamma_{+,n}(h_n)] &= E_{P_n}\left[\frac{1}{nh_n}\sum_{i=1}^n 1\{Z_i \geq 0\}K(Z_i/h_n)r(Z_i/h_n)r(Z_i/h_n)'\right] \\ &= \frac{1}{h_n}\int_0^\infty K(z/h_n)r(z/h_n)r(z/h_n)'f_{P_n}(z)dz \\ &= \int_0^\infty K(u)r(u)r(u)'f_{P_n}(uh_n) \equiv \tilde{\Gamma}_{+,n}(h_n) \ . \end{aligned}$$

Further, since $h_n < \tilde{\kappa}$ for large enough n, we have that $\tilde{L} \leq f_{P_n}(z) \leq \tilde{U}$ by Assumption 4.1, which implies that

$$\Gamma_L \equiv \tilde{L}\int_0^\infty K(u)r(u)r(u)'du \leq \tilde{\Gamma}_{+,n}(h_n) \leq \tilde{U}\int_0^\infty K(u)r(u)r(u)'du \equiv \Gamma_U \ ,$$

and that

$$\begin{aligned} E_{P_n^n}[|\Gamma_{+,n}(h_n) - E_{P_n}[\Gamma_{+,n}(h_n)]|^2] &\leq \frac{1}{h_n^2} E_{P_n}\left[\left|1\{Z_i \geq 0\} K(Z_i/h_n) r(Z_i/h_n) r(Z_i/h_n)'\right|^2\right] \\ &= \frac{1}{nh_n} \int_0^\infty K(u)^2 |r(u)|^4 f_{P_n}(uh_n) du \\ &\leq \frac{\tilde{U}}{nh_n} \int_0^\infty K(u)^2 |r(u)|^4 du \\ &= O(n^{-1}h_n^{-1}) = o(1) \ . \end{aligned}$$

It then follows by Markov's Inequality that $\Gamma_{+,n}(h_n) = \tilde{\Gamma}_{+,n}(h_n) + o_{P_n}(1)$. Analogously, closely following Calonico et al. (2014c, Lemma S.A.1), we can show Lemma C.0.1(ii)-(iii).

■

Lemma C.0.2. *Let* $\mathbf{Q}$ *be defined as in* (C.3), $\mathbf{P}$ *be as in* (C.4) *and* $P_n \in \mathbf{P}$ *for all* $n \geq 1$. *If* $nh_n \to \infty$ *and* $h_n \to 0$, *then*

(i) $E_{P_n^n}[\hat{\mu}_{+,n}|Z^{(n)}] = \mu_+(P_n) + O_{P_n}(h_n^3)$.

(ii) $V_{P_n^n}[\hat{\mu}_{+,n}|Z^{(n)}] = n^{-1} e' \Gamma_{+,n}^{-1}(h_n) \Psi_{+,n}(h_n, P_n) \Gamma_{+,n}^{-1}(h_n) e \equiv V_{+,n}(h_n, P_n)$.

(iii) $(V_{+,n}(h_n, P_n))^{-1/2} \left(\hat{\mu}_{+,n} - E_{P_n^n}[\hat{\mu}_{+,n}|Z^{(n)}]\right) \xrightarrow{d} \mathcal{N}(0,1)$.

Proof. For (i), by taking the conditional on $Z^{(n)}$ expectation, we have

$$E_{P_n^n}[\hat{\mu}_{+,n}|Z^{(n)}] = e' H(h_n) \Gamma_{+,n}^{-1}(h_n) Z_n(h_n)' W_{+,n}(h_n) E_{P_n^n}[\mathbf{Y}_n|Z^{(n)}]/n \ .$$

Further, as $h_n < \tilde{\kappa}$ for large enough n, we have by the required differentiability in Assumption 4.1 and a Taylor expansion for $0 < Z < h_n$ that

$$E_{P_n}[Y|Z] = \mu_+(P_n) + Z\mu_+^1(P_n) + (Z/2)^2 \mu_+^2(P_n) + O_{P_n}(h_n^3) \ .$$

It then follows from Lemma C.0.1 and the previous two expressions that

$$E_{P_n^n}[\hat{\mu}_+|Z^{(n)}] = \mu_+(P_n) + O_{P_n}(h_n^3) \ .$$

For (ii), a simple calculation gives us

$$\begin{aligned} V_{P_n^n}[\hat{\mu}_+(h_n)|Z^{(n)}] &= e'H(h_n)\Gamma_{+,n}^{-1}(h_n)Z_n(H_n)'W_{+,n}(h_n)\Sigma_n(P_n)W_{+,n}(h_n)Z_n(h_n)\Gamma_{+,n}^{-1}(h_n)H(h_n)e/n^2 \\ &= n^{-1}e'\Gamma_{+,n}^{-1}(h_n)\Psi_{+,n}(h_n,P_n)\Gamma_{+,n}^{-1}(h_n)e \equiv V_{+,n}(h_n,P_n) \ . \end{aligned}$$

For (iii), first note that from Lemma C.0.1 we have $V_{+,n}(h_n,P_n) = \tilde{V}_{+,n}(h_n) + o_{P_n}(1)$, where

$$\tilde{V}_{+,n}(h_n) = (nh_n)^{-1}e'\tilde{\Gamma}_{+,n}^{-1}(h_n)\tilde{\Psi}_{+,n}(h_n)\tilde{\Gamma}_{+,n}^{-1}(h_n)e \ .$$

Then rewrite as follows

(C.5)

$$\frac{\hat{\mu}_{+,n} - E_{P_n^n}[\hat{\mu}_{+,n}|Z^{(n)}]}{\sqrt{V_{+,n}(h_n,P_n)}} = \left(\frac{\tilde{V}_{+,n}(h_n,P_n)}{V_{+,n}(h_n,P_n)}\right)^{1/2}\left(\tilde{V}_{+,n}(h_n)\right)^{-1/2} e'\Gamma_{+,n}^{-1}(h_n)\tilde{\Gamma}_{+,n}(h_n)\tilde{A}_n^{1/2}\xi_n \ ,$$

where

$$\begin{aligned} \xi_n &= \sum_{i=1}^n \omega_{n,i}\epsilon_{n,i} \ , \\ \epsilon_{n,i} &= Y_i - E_{P_n}[Y_i|Z_i] \ , \\ \tilde{A}_n &= (nh_n)^{-1}\tilde{\Gamma}_{+,n}^{-1}(h_n)\tilde{\Psi}_{+,n}(h_n)\tilde{\Gamma}_{+,n}^{-1}(h_n) \ , \text{ and} \\ \omega_{n,i} &= \tilde{A}_n^{-1/2}\tilde{\Gamma}_{+,n}^{-1}(h_n)K_{h_n}(Z_i/h_n)r(Z_i/h_n)/n \ . \end{aligned}$$

Next note that for any $a \in \mathbf{R}^3$ we have that $\{a'\omega_{n,i}\epsilon_{n,i} : 1 \leq i \leq n\}$ is a triangular array of independent random variables with $E_{P_n^n}[a'\xi_n] = 0$ and $V_{P_n^n}[a'\xi_n] = a'a$. Further, this triangular array satisfies the Lindeberg condition. To see why, first note that by Lemma C.0.1 we have

$$|\tilde{A}_n| \geq (nh_n)^{-1}|\tilde{A}_L| \ , \tag{C.6}$$

for some value $\tilde{A}_L \in \mathbf{R}$, which is uniform in $\mathbf{P}$. We then have by Lemma C.0.1 and a change of variables that

$$\begin{aligned}
\sum_{1=1}^{n} E_{P_n}[|a'\omega_{n,i}\epsilon_i|^4] &\precsim (nh_n)^2 \sum_{1=1}^{n} E_{P_n}\left[|a'K_{h_n}(Z/h_n)r(Z/h_n)/n|^4\right] \\
&\precsim (nh_n)^2 n^{-3} h_n^{-4} \int_0^\infty |a'K(z/h_n)r(z/h_n)|^4 f_{P_n}(z)dz \\
&\precsim (nh_n)^2 n^{-3} h_n^{-3} = O\left((nh_n)^{-1}\right) = o(1)
\end{aligned}$$

and hence, using the Lindeberg-Feller CLT, we have that $a'\xi_n \xrightarrow{d} \mathcal{N}(0, a'a)$ as $n \to \infty$. Since this holds for any $a \in \mathbf{R}^3$, the Cramér-Wold theorem implies that $\xi_n \xrightarrow{d} \mathcal{N}(0, I_3)$ as $n \to \infty$, where I_3 denotes the identity matrix of size three. Furthermore, analogous to $V_+(h_n, P_n) = \tilde{V}_+(h_n) + o_{P_n}(1)$, we can show that

$$\frac{V_{+,n}(h_n, P_n)}{\tilde{V}_{+,n}(h_n)} = 1 + o_{P_n}(1) \ . \tag{C.7}$$

Further, by Lemma C.0.1 we have that

$$\Gamma_{+,n}^{-1}(h_n)\tilde{\Gamma}_{+,n}(h_n) = I_3 + o_{P_n}(1) \ . \tag{C.8}$$

Substituting the above results in (C.5) concludes the proof. ■

C.0.3. Proof of Theorem 4.2

Here we show only that

$$\frac{\hat{\mu}_{+,n} - \mu_+(P_n)}{\sqrt{\hat{V}_{+,n}}} \xrightarrow{d} \mathcal{N}(0,1) \ ,$$

since under similar arguments it will follow that

$$\frac{\hat{\mu}_{n,-} - \mu_-(P_n)}{\sqrt{\hat{V}_{n,-}}} \xrightarrow{d} \mathcal{N}(0,1) \ ,$$

and then due to independence we can conclude that $T_n^{CCT} \xrightarrow{d} \mathcal{N}(0,1)$ as $n \to \infty$. To this end, first rewrite

$$\frac{\hat{\mu}_{+,n} - \mu_+(P_n)}{\sqrt{\hat{V}_{+,n}}} = \frac{\hat{\mu}_{+,n} - \mu_+(P_n)}{\sqrt{V_{+,n}(h_n, P_n)}} \cdot \sqrt{\frac{V_{+,n}(h_n, P_n)}{\hat{V}_{+,n}}} \ .$$

Step 1. We show that

$$\frac{\hat{\mu}_{+,n} - \mu_+(P_n)}{\sqrt{V_{+,n}(h_n, P_n)}} \xrightarrow{d} \mathcal{N}(0,1) \ . \tag{C.9}$$

To begin, first rewrite the above as

$$\frac{\hat{\mu}_{+,n} - E_{P_n^n}[\hat{\mu}_{+,n}|Z^{(n)}]}{\sqrt{V_{+,n}(h_n, P_n)}} + \left(\frac{\tilde{V}_{+,n}(h_n)}{V_{+,n}(h_n, P_n)}\right)^{1/2} \frac{E_{P_n^n}[\hat{\mu}_{+,n}|Z^{(n)}] - \mu_+(P_n)}{\sqrt{\tilde{V}_{+,n}(h_n)}} \ .$$

In Lemma C.0.2 (iii), we showed that

$$\frac{\hat{\mu}_{+,n} - E_{P_n^n}[\hat{\mu}_{+,n}|Z^{(n)}]}{\sqrt{V_{+,n}(h_n, P_n)}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{and} \quad \frac{\tilde{V}_{+,n}(h_n)}{V_{+,n}(h_n, P_n)} = 1 + o_{P_n}(1) \ .$$

It then remains to show that

$$\frac{E_{P_n^n}[\hat{\mu}_{+,n}|Z^{(n)}] - \mu_+(P_n)}{\sqrt{\tilde{V}_{+,n}(h_n)}} = o_{P_n}(1) \ ,$$

to conclude. To this end, note that by Lemma C.0.2 and (C.6), it follows that

$$\frac{\left|E_{P_n^n}[\hat{\mu}_{+,n}|Z^{(n)}] - \mu_+(P_n)\right|}{\sqrt{\tilde{V}_{+,n}(h_n)}} = O\left((nh_n)^{1/2}\right) O_{P_n}(h_n^3) = O_{P_n}\left((nh_n^7)^{1/2}\right) = o_{P_n}(1)$$

as $h_n \to 0$, $nh_n \to \infty$ and $nh_n^7 \to 0$.

Step 2. We show that

$$\frac{V_{+,n}(h_n, P_n)}{\hat{V}_{+,n}} = 1 + o_{P_n}(1) \ . \tag{C.10}$$

To begin note that

(C.11)

$$nh_n\left(V_{+,n}(h_n, P_n) - \hat{V}_{+,n}\right) = e'\Gamma_{+,n}^{-1}(h_n) \cdot h_n\left(\Psi_{+,n}(h_n, P_n) - \hat{\Psi}_{+,n}(h_n)\right) \cdot \Gamma_{+,n}^{-1}(h_n)e \ ,$$

where

(C.12)

$$h_n\left(\Psi_{+,n}(h_n, P_n) - \hat{\Psi}_{+,n}(h_n)\right) = h_n Z_n(h_n)' W_{+,n}(h_n)\left(\Sigma_n(P_n) - \hat{\Sigma}_n\right) W_{+,n}(h_n) Z_n(h_n)/n \ ,$$

and

$$\hat{\Sigma}_{+,n} = \text{diag}(\hat{\epsilon}^2_{+,n,1}, \ldots, \hat{\epsilon}^2_{+,n,n}) \ , \tag{C.13}$$

such that $\hat{\epsilon}_{+,n,i} = Y_i - \hat{\mu}_{+,n}$. Further, note that by construction, we can write

$$Y_i = \mu(Z_i, P_n) + \epsilon_{n,i} \ , \tag{C.14}$$

such that $E_{P_n}[\epsilon_{n,i}] = 0$ and $Var_{P_n}[\epsilon_{n,i}|Z = z] = \sigma^2(z, P_n)$. This in turn implies

$$\hat{\epsilon}_{+,n,i} = \epsilon_{n,i} + \mu(Z_i, P_n) - \mu_+(P_n) + \mu_+(P_n) - \hat{\mu}_{+,n} \ . \tag{C.15}$$

We can then expand (C.12) to get the following

$$\begin{aligned} h_n \left(\Psi_{+,n}(h_n, P_n) - \hat{\Psi}_{+,n}(h_n) \right) = & \underbrace{h_n \sum_{i=1}^{n} 1\{Z_i \geq 0\} (\sigma^2(Z_i, P_n) - \epsilon^2_{n,i}) K_{h_n}(Z_i)^2 r(Z_i/h_n) r(Z_i/h_n)'/n}_{\equiv B_{1,n} \ , \text{ (a)}} \\ & \underbrace{- h_n \sum_{i=1}^{n} 1\{Z_i \geq 0\} (\mu(Z_i, P_n) - \hat{\mu}_{+,n})^2 K_{h_n}(Z_i)^2 r(Z_i/h_n) r(Z_i/h_n)'/n}_{\equiv B_{2,n} \ , \text{ (b)}} \\ & \underbrace{+ 2\, h_n \sum_{i=1}^{n} 1\{Z_i \geq 0\} \epsilon_{n,i} (\mu(Z_i, P_n) - \hat{\mu}_{+,n}) K_{h_n}(Z_i)^2 r(Z_i/h_n) r(Z_i/h_n)'/n}_{\equiv B_{3,n} \ , \text{ (c)}} \ . \end{aligned}$$

For quantity (a), since Assumption 2.1 (i), Assumption 2.1 (ii) and Assumption 2.1 (iv) are satisfied with the required uniform constants, we have by a change of variables that

$$\begin{aligned} E_{P_n}\left[|B_{1,n}|^2 \right] &\precsim (nh_n)^{-1} \int_0^\infty K(u)^4 |r(u)|^4 du \\ &= O((nh_n)^{-1}) = o(1) \ , \end{aligned}$$

which implies by Markov's Inequality that $B_{n,1} = o_{P_n}(1)$. For quantity (b), note that first we can rewrite it as

$$\begin{aligned} B_{n,2} = & \underbrace{h_n \sum_{i=1}^{n} 1\{Z_i \geq 0\}(\mu(Z_i, P_n) - \mu_+(P_n))^2 K_{h_n}(Z_i)^2 r(Z_i/h_n) r(Z_i/h_n)'/n}_{\equiv B_{n,21}} \\ & + (\mu_+(P_n) - \hat{\mu}_{+,n})^2 \cdot \underbrace{h_n \sum_{i=1}^{n} 1\{Z_i \geq 0\} K_{h_n}(Z_i)^2 r(Z_i/h_n) r(Z_i/h_n)'/n}_{\equiv B_{n,22}} \\ & + 2(\mu_+(P_n) - \hat{\mu}_{+,n}) \cdot \underbrace{h_n \sum_{i=1}^{n} 1\{Z_i \geq 0\}(\mu(Z_i, P_n) - \mu_+(P_n)) K_{h_n}(Z_i)^2 r(Z_i/h_n) r(Z_i/h_n)'/n}_{\equiv B_{n,23}} , \end{aligned}$$

Next, since Assumption 2.1 (i) and Assumption 2.1 (iii) are satisfied with the required uniform constants, we have by a Taylor approximation and a change of variables that

$$\begin{aligned} E_{P_n}[|B_{n,21}|^2] &\precsim n^{-1} h_n^3 \int_0^\infty K(u)^4 |r(u)|^4 du \\ &= O(n^{-1} h_n^3) = o(1) , \end{aligned}$$

which implies by Markov's inequality that $B_{n,21} = o_{P_n}(1)$. Further, since Assumption 2.1 (i) is satisfied with the required uniform constants, we have by a change of variables that

$$\begin{aligned} E_{P_n}[|B_{n,22}|^2] &\precsim (nh_n)^{-1} \int_0^\infty K(u)^4 |r(u)|^4 du \\ &= O((nh_n)^{-1}) = o(1) , \end{aligned}$$

which implies by Markov's inequality that $B_{n,22} = o_{P_n}(1)$. Finally, since Assumption 2.1 (i) and Assumption 2.1 (iii) are satisfied with the required uniform constants, we have by

a Taylor approximation and a change of variables that

$$E_{P_n}\left[|B_{n,23}|^2\right] \precsim (n)^{-1}h_n \int_0^\infty K(u)^4 |r(u)|^4 du$$
$$= O(n^{-1}h_n) = o(1) \ ,$$

which implies by Markov's inequality that $B_{n,23} = o_{P_n}(1)$. Since $\mu_+(P_n) - \hat{\mu}_{+,n} = o_{P_n}(1)$ by (C.9), we can conclude for quantity (b) that $B_{n,2} = o_{P_n}(1)$. For quantity (c), using analogous arguments, we can conclude that $B_{n,3} = o_{P_n}(1)$, and hence

$$h_n \left(\Psi_{+,n}(h_n, P_n) - \hat{\Psi}_{+,n}(h_n)\right) = o_{P_n}(1) \ . \tag{C.16}$$

In addition, since from Lemma C.0.1 we have that $\Gamma^{-1}_{+,n}(h_n) = \tilde{\Gamma}^{-1}_{+,n}(h_n)$, it then follows that

$$nh_n \left(V_{+,n}(h_n, P_n) - \hat{V}_{+,n}\right) = o_{P_n}(1) \ . \tag{C.17}$$

To conclude, first rewrite (C.17) as

$$\frac{V_{+,n}(h_n, P_n) - \hat{V}_{+,n}}{\tilde{V}_{+,n}(h_n)} = o_{P_n}(1) \ ,$$

and our result then follows from (C.7).