Extrapolation and Bandwidth Choice in the Regression Discontinuity Design

by

Miikka Rokkanen

M.Sc. Economics, University of Jyväskylä (2008)
B.Sc. Economics, University of Jyväskylä (2008)

Submitted to the Department of Economics
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2014

Signature redacted

Author .....
Department of Economics
May 15, 2014

Signature redacted

Certified by ....
Joshua Angrist
Ford Professor of Economics
Thesis Supervisor

Signature redacted

Certified by ....
Parag Pathak
Associate Professor of Economics
Thesis Supervisor

Signature redacted

Accepted by ....
Michael Greenstone
3M Professor of Economics
Chairman, Departmental Committee on Graduate Studies

Extrapolation and Bandwidth Choice in the Regression Discontinuity Design

by

Miikka Rokkanen

Submitted to the Department of Economics
on May 15, 2014, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

This thesis consists of three methodological contributions to the literature on the regression discontinuity (RD) design. The first two chapters develop approaches to the extrapolation of treatment effects away from the cutoff in RD and use them to study the achievement effects of attending selective public schools, known as exam schools, in Boston. The third chapter develops an adaptive bandwidth choice algorithm for local polynomial regression-based RD estimators.

The first chapter develops a latent factor-based approach to RD extrapolation that is then used to estimate effects of exam school attendance for inframarginal 7th grade applicants. Achievement gains from Boston exam schools are larger for applicants with lower English and Math abilities. I also use the model to predict the effects of introducing either minority or socioeconomic preferences in exam school admissions. Affirmative action has modest average effects on achievement, while increasing the achievement of the applicants who gain access to exam schools as a result.

The second chapter, written jointly with Joshua Angrist, develops a covariate-based approach to RD extrapolation that is then used to estimate effects of exam school attendance for inframarginal 9th grade applicants. The estimates suggest that the causal effects of exam school attendance for applicants with running variable values well away from admissions cutoffs differ little from those for applicants with values that put them on the margin of acceptance.

The third chapter develops an adaptive bandwidth choice algorithm for local polynomial regression-based RD estimators. The algorithm allows for different choices for the order of polynomial and kernel function. In addition, the algorithm automatically takes into account the inclusion of additional covariates as well as alternative assumptions on the variance-covariance structure of the error terms. I show that the algorithm produces a consistent estimator of the asymptotically optimal bandwidth and that the resulting regression discontinuity estimator satisfies the asymptotic optimality criterion of Li (1987). Finally, I provide Monte Carlo evidence suggesting that the proposed algorithm also performs well in finite samples.

Thesis Supervisor: Joshua Angrist
Title: Ford Professor of Economics

Thesis Supervisor: Parag Pathak
Title: Associate Professor of Economics

Contents

Acknowledgements

I would like to thank my thesis supervisors Joshua Angrist and Parag Pathak as well as Victor Chernozhukov and Whitney Newey for their guidance and support. I would like to the faculty and students at MIT for creating such an amazing research environment. I would like to thank the staff at MIT for all their help with various practical matters. I would like to thank the seminar participants and countless other people I have discussed about my research with over the years for their insightful comments and suggestions. I would like to thank Matti Sarvimäki and Roope Uusitalo for everything they have done for me both before and during my graduate studies. I would like to thank Aalto University, Government Institute for Economic Research and Helsinki Center of Economic Research for their hospitality during my visits to Finland. I would like to thank Yrjö Jahnsson Foundation for financial support. Last but not least, I would like to thank my family and friends for making my life so enjoyable.

List of Tables

List of Figures

Chapter 1

Exam Schools, Ability, and the Effects of Affirmative Action: Latent Factor Extrapolation in the Regression Discontinuity Design

1.1 Introduction

Regression Discontinuity (RD) methods identify treatment effects for individuals at the cutoff value determining treatment assignment under relatively mild assumptions (Hahn, Todd, and van der Klaauw, 2001; Frandsen, Frolich, and Melly, 2012).[1] Without stronger assumptions, however, nothing can be said about treatment effects for individuals away from the cutoff. Such effects may be valuable for predicting the effects of policies that change treatment assignments of a broader group. An important example of this are affirmative action policies that change cutoffs substantially.

Motivated by affirmative action considerations, this paper develops a strategy for the identification and estimation of causal effects for inframarginal applicants to Boston's selective high schools, known as exam schools. The exam schools, spanning grades 7-12, are seen as the flagship of the Boston Public Schools (BPS) system. They offer higher-achieving peers and an advanced curriculum. Admissions to these schools are based on Grade Point Average (GPA) and the Independent School Entrance Exam (ISEE). The RD design generated by exam school admissions nonparametrically identifies causal effects of exam school attendance for marginal applicants at admissions cutoffs. Abdulkadiroglu, Angrist, and Pathak (2014) use this strategy

[1] Cook (2008) provides an extensive treatment of the history of RD. See also the surveys by Imbens and Lemieux (2008), van der Klaauw (2008), Imbens and Wooldridge (2009), Lee and Lemieux (2010), and DiNardo and Lee (2011).

and find little evidence of effects for these applicants.[2] Other applicants, however, may benefit or suffer as a consequence of exam school attendance.

Treatment effects away from RD cutoffs are especially important for discussions of affirmative action at exam schools. Boston exam schools have played an important role in the history of attempts to ameliorate racial imbalances in Boston. In 1974 a federal court ruling introduced the use of minority preferences in Boston exam school admissions as part of a city-wide desegregation plan. Court challenges later led the Boston school district to drop racial preferences. Similarly, Chicago switched from minority to socioeconomic preferences in exam school admissions following a federal court ruling in 2009.[3]

This paper develops a latent factor-based approach to the identification and estimation of treatment effects away from the cutoff. I assume that the source of omitted variables bias in an RD design can be modeled using latent factors. The running variable is one of a number of noisy measures of these factors. Assuming other noisy measures are available, causal effects for all values of the running variable are nonparametrically identified.[4] In related work on the same problem, Angrist and Rokkanen (2013) postulate a strong conditional independence assumption that identifies causal effects away from RD cutoffs. The framework developed here relies on weaker assumptions and is likely to find wider application.[5]

I use this framework to estimate causal effects of exam school attendance for the full population of applicants. These estimates suggest that the achievement gains from exam school attendance are larger among applicants with lower baseline measures of ability. I also use the latent factor framework to simulate effects of introducing either minority or socioeconomic preferences in Boston exam school admissions. These reforms change the admissions cutoffs faced by diffent applicant groups and affect the exam school assignment of 27-35% of applicants. The simulations suggest that the reforms boost achievement among applicants. These effects are largely driven by achievement gains experienced by lower-achieving applicants who gain access to exam schools as a result.

In developing the latent factor-based approach to RD extrapolation I build on the literatures on measurement error models (Kotlarski, 1967; Hu and Schennach, 2008; Evdokimov and White, 2012) and (semi-)nonparametric instrumental variable models (Newey and Powell, 2003; Darolles, Fan, Florens, and Renault, 2011).[6] Latent factor models have a long tradition in economics (Aigner, Hsiao, Kapteyn, and Wansbeek, 1984). In the program evaluation literature, for instance, latent factor models have been used to identify the joint distribution of potential outcomes (Carneiro, Hansen, and Heckman, 2001, 2003; Aakvik, Heckman, and Vytlacil, 2005; Cunha, Heckman, and Navarro, 2005; Battistin, Lamarche, and Rettore, 2013), time-varying

[2]Dobbie and Fryer (2013) find similar results in an RD study of New York City exam schools.

[3]The use of affirmative action in exam school admissions is a highly contentious issue also in New York City where a federal complaint was filed in 2012 against the purely achievement-based exam school admissions processdue to disproportionately low minority shares at these schools.

[4]This is similar to ideas put forth by Lee (2008), Lee and Lemieux (2010), DiNardo and Lee (2011), and Bloom (2012). However, this is the first paper that discusses how this framework can be used in RD extrapolation.

[5]For other approaches to RD extrapolation, see Angrist and Pischke (2009), Jackson (2010), DiNardo and Lee (2011), Dong and Lewbel (2013), Cook and Wing (2013), and Bargain and Doorley (2013).

[6]See also the surveys by Hausman (2001), Blundell and Powell (2003), Chen, Hong, and Nekipelov (2011), and Horowitz (2011) as well as the references therein.

treatment effects (Cooley Fruehwirth, Navarro, and Takahashi, 2011), and distributional treatment effects (Bonhomme and Sauder, 2011).

The educational consequences of affirmative action have mostly been studied in post-secondary schools with a focus on application and enrollment margins. Several papers have studied affirmative action bans in California and Texas as well as the introduction of the Texas 10% plan (Long, 2004; Card and Krueger, 2005; Dickson, 2006; Andrews, Ranchhod, and Sathy, 2010; Cortes, 2010; Antonovics and Backes, 2013, forthcoming). Howell (2010) uses a structural model to simulate the effects of a nation-wide elimination of affirmative action in college admissions, and Hinrichs (2012) studies the effects of various affirmative action bans around the United States. Only a few studies have looked at the effects of affirmative action in selective school admissions on later outcomes (Arcidiacono, 2005; Rothstein and Yoon, 2008; Bertrand, Hanna, and Mullainathan, 2010; Francis and Tannuri-Pianto, 2012).

The rest of the paper is organized as follows. The next section outlines the econometric framework. Section 3 discusses extensions of this approach to fuzzy RD and settings with multiple latent factors. Section 4 discusses identification and estimation of the latent factor model in the Boston exam school setting. Section 5 reports latent factor estimates. Section 6 uses the model to analyse effects of affirmative action. Section 7 concludes.

1.2 Latent Factor Modeling in a Sharp RD Design

1.2.1 Framework

Suppose one is interested in the causal effect of a binary treatment $D \in \{0, 1\}$ on an outcome $Y \in \mathcal{Y}$ that can be either discrete or continuous. Each individual is associated with two potential outcomes: $Y(0)$ is the outcome of an individual if she is not exposed to the treatment ($D = 0$), and $Y(1)$ is the outcome of an individual if she is exposed to the treatment ($D = 1$). The observed outcome of an individual is

$$Y = Y(0)(1 - D) + Y(1) D.$$

In a sharp Regression Discontinuity (RD) design the treatment assignment is fully determined by whether the value of a continuous covariate $R \in \mathcal{R}$, often called the running variable, lies above or below a known cutoff c.[7] That is, the treatment assignment is given by

$$D = 1(R \geq c).$$

I ignore the presence of additional covariates to simplify the notation. It is possible to generalize all the results to allow for additional covariates by conditioning on them throughout.

[7] In a fuzzy RD the treatment assignment is only partially determined by the running variable. I discuss this extension in Section 1.3.1.

The sharp RD design allows one to nonparametrically identify the Average Treatment Effect (ATE) at the cutoff, $E[Y(1) - Y(0) \mid R = c]$, under the conditions listed in Assumption A. Assumption A.1 restricts the marginal density of R to be strictly positive in a neighborhood of the cutoff c. Assumption A.2 restricts the conditional cumulative distribution functions of both $Y(0)$ and $Y(1)$ given R to be continuous in R at the cutoff c.[8] Finally, Assumption A.3 requires that the conditional expectations of $Y(0)$ and $Y(1)$ exist at the cutoff c. Under these assumptions, the Average Treatment Effect at the cutoff is given by the discontinuity in the conditional expectation function of Y given R at the cutoff, as shown in Lemma 1 (Hahn, Todd, and van der Klaauw, 2001).

Assumption A.

1. $f_R(r) > 0$ *in a neighborhood around* c.
2. $F_{Y(0)|R}(y \mid r)$ *and* $F_{Y(1)|R}(y \mid r)$ *are continuous in* r *at* c *for all* $y \in \mathcal{Y}$.
3. $E[|Y(0)| \mid R = c], E[|Y(1)| \mid R = c] < \infty$.

Lemma 1. *(Hahn, Todd, and van der Klaauw, 2001) Suppose Assumption A holds. Then*

$$E[Y(1) - Y(0) \mid R = c] = \lim_{\delta \downarrow 0} \{E[Y \mid R = c + \delta] - E[Y \mid R = c - \delta]\}.$$

Lemma 1 illustrates the power of sharp RD as it nonparametrically identifies the Average Treatment Effect at the cutoff under relatively mild assumptions. However, there is not much one can say about the Average Treatment Effect away from the cutoff without stronger assumptions. Figure 1-1 illustrates this extrapolation problem. To the left of the cutoff one observes $Y(0)$ as these individuals are not assigned to the treatment whereas to the right of the cutoff one observes $Y(1)$ as these individuals are assigned to the treatment. The relevant counterfactual outcomes are instead unobservable.

To motivate the importance of extrapolation away from the cutoff, suppose one wanted to know the Average Treatment Effect for individuals with $R = r_0$ to the left of the cutoff. For these individuals one observes $E[Y(0) \mid R = r_0]$, but the counterfactual $E[Y(1) \mid R = r_0]$ is unobservable. Similarly, suppose one wanted to know the Average Treatment Effect for individuals with $R = r_1$ to the right of the cutoff. For these individuals one observes $E[Y(1) \mid R = r_1]$, but the counterfactual $E[Y(0) \mid R = r_1]$ is unobservable.

In this paper I develop a latent factor-based solution to the extrapolation problem. Consider a setting in which R is a function of a latent factor θ and disturbance ν_R:

$$R = g_R(\theta, \nu_R)$$

where g_R is an unknown function, and both θ and ν_R are potentially multidimensional. Suppose, for instance, that R is an entrance exam score used in admissions to a selective school. Then, it is natural to interpret R

[8] Continuity of the conditional expectation functions of the potential outcomes is enough for Lemma 1. However, continuity of the conditional cumulative distribution functions of the potential outcomes allows one to also identify distributional treatment effects (Frandsen, Frolich, and Melly, 2012).

as a noisy measure of an applicant's academic ability.

Figure 1-2 illustrates the latent factor framework when both θ and ν_R are scalars and $R = \theta + \nu_R$. Consider two types of individuals with low and high levels of θ, θ^{low} and θ^{high}. Furthermore, suppose that $\theta^{low} < c$ and $\theta^{high} > c$. Then, if there was no noise in R, individuals with $\theta = \theta^{low}$ would not receive the treatment whereas individuals with $\theta = \theta^{high}$ would receive the treatment. However, because of the noise in R some of the individuals with $\theta = \theta^{low}$ end up to the right of the cutoff, and similarly some of the individuals with $\theta = \theta^{high}$ end up to the left of the cutoff. Thus, both types of individuals are observed with and without the treatment.

I assume that the potential outcomes $Y(0)$ and $Y(1)$ are conditionally independent of R given θ, as stated in Assumption B. This means that any dependence between $(Y(0), Y(1))$ and R is solely due to two factors: the dependence of $Y(0)$ and $Y(1)$ on θ and the dependence of R on θ.

Assumption B. $(Y(0), Y(1)) \perp\!\!\!\perp R \mid \theta$.

Lemma 2. *Suppose that Assumption B holds. Then*

$$E[Y(1) - Y(0) \mid R = r] = E\{E[Y(1) - Y(0) \mid \theta] \mid R = r\}$$

for all $r \in \mathcal{R}$.

Lemma 2 highlights the key implication of Assumption B. Under this assumption, the conditional Average Treatment Effect given $R = r$, $E[Y(1) - Y(0) \mid R = r]$, depends on two objects: the latent conditional Average Treatment Effect given θ, $E[Y(1) - Y(0) \mid \theta]$, and the conditional distribution of θ given R, $f_{\theta|R}$.[9] Thus, the identification of the Average Treatment Effect away from the cutoff depends on one's ability two identify these two objects. In the selective school admissions example Assumption B means that while $Y(0)$ and $Y(1)$ may depend on an applicant's academic ability, they do not depend on the noise in the entrance exam score.

Figure 1-5 illustrates this by considering again the identification of the Average Treatment Effect for individuals with $R = r_0$ to the left of the cutoff and for individuals with $R = r_1$ to the right of the cutoff. As discussed above, the shap RD design allows one to observe $E[Y(0) \mid R = r_0]$ and $E[Y(1) \mid R = r_1]$, and the extrapolation problem arises from the unobservability of $E[Y(1) \mid R = r_0]$ and $E[Y(0) \mid R = r_1]$. Suppose the conditional expectation functions of $Y(0)$ and $Y(1)$ given θ, $E[Y(0) \mid \theta]$ and $E[Y(1) \mid \theta]$, depicted in Figure 1-3, are known. In addition, suppose the conditional densities of θ given $R = r_0$ and $R = r_1$, $f_{\theta|R}(\theta \mid r_0)$ and $f_{\theta|R}(\theta \mid r_1)$, depicted in Figure 1-4, are known. Then, under Assumption B, the

[9]For all the results in this section it is enough to assume that $Y(0)$ and $Y(1)$ are conditionally mean independent of R given θ. However, the full conditional independence assumption stated here allows also for the extrapolation of distributional treatment effects away from the cutoff. I do not discuss this in detail as it is a straightforward extension of the results presented below.

counterfactuals $E[Y(1) \mid R = r_0]$ and $E[Y(0) \mid R = r_1]$ are given by

$$\begin{aligned} E[Y(1) \mid R = r_0] &= E\{E[Y(1) \mid \theta] \mid R = r_0\} \\ E[Y(0) \mid R = r_1] &= E\{E[Y(0) \mid \theta] \mid R = r_1\}. \end{aligned}$$

There is only one remaining issue: how does one identify the latent conditional Average Treatment Effect given θ, $E[Y(1) - Y(0) \mid \theta]$, and the conditional distribution of θ given R, $f_{\theta|R}$? If θ was observable, these objects could be identified using the covariate-based approach developed by Angrist and Rokkanen (2013). However, here θ is an unobservable latent factor which complicates the identification of $E[Y(1) - Y(0) \mid \theta]$ and $f_{\theta|R}$.[10] To achieve identification, I rely on the availability of multiple noisy measures of θ. To simplify the discussion, I consider in this section a setting in which θ is unidimensional. I discuss an extension of the approach to settings with multidimensional latent factors in Section 1.3.2.

I assume that the data contains three noisy measures of θ, denoted by M_1, M_2, and M_3:

$$\begin{aligned} M_1 &= g_{M_1}(\theta, \nu_{M_1}) \\ M_2 &= g_{M_2}(\theta, \nu_{M_2}) \\ M_3 &= g_{M_3}(\theta, \nu_{M_3}) \end{aligned}$$

where g_{M_1}, g_{M_2}, and g_{M_3} are unknown functions, and ν_{M_1}, ν_{M_2}, and ν_{M_3} are potentially multidimensional disturbances. I focus on a setting in which is R is a deterministic function of at least one or potentially many of these measures, but it is possible to consider a more general setting that allows the relationship between R and M to be stochastic. Going back to the selective school example considered above, one might think of M_1 as the entrance exam score whereas M_2 and M_3 might be two pre-application baseline test scores.

I require θ, M_1, and M_2 to be continuous but allow M_3 to be either continuous or discrete; even a binary M_3 is sufficient. I denote the supports of θ, M_1, M_2, and M_3 by Θ, $\mathcal{M}_1$, $\mathcal{M}_2$, and $\mathcal{M}_3$. I occasionally also use the notation $M = (M_1, M_2, M_3)$ and $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{M}_3$. I leave the properties of the latent factor θ, the unknown functions g_{M_1}, g_{M_2}, and g_{M_3} as well as the disturbances ν_{M_1}, ν_{M_2}, and ν_{M_3} unspecified for now. I return to them below when discussing alternative sets of assumptions allowing for the identification of the measurement model.

1.2.2 Parametric Illustration

To provide a benchmark for the discussion about nonparametric identification of the latent factor model, I begin by considering the identification of a simple parametric model. I assume linearity and normality in the measurement models for M_1 and M_2 but leave the measurement model for M_3 flexible. In addition, I

[10] The covariate-based approach by Angrist and Rokkanen (2013) can be in certain cases used for extrapolation even if the conditional independence assumption holds only for a latent factor. For this assumption to work, one needs to assume that the running variable contains no additional information about the latent factor once one conditions on a set of covariates.

assume linearity in the latent outcome models $E[Y(0) \mid \theta]$ and $E[Y(1) \mid \theta]$.

The measurement model takes the following form:

$$\begin{aligned} M_1 &= \theta + \nu_{M_1} \\ M_2 &= \mu_{M_2} + \lambda_{M_2}\theta + \nu_{M_2} \\ M_3 &= g(\theta, \nu_{M_3}) \end{aligned}$$

where $\lambda_{M_2}, Cov(\theta, M_3) \neq 0$, and

$$\begin{bmatrix} \theta \\ \nu_{M_1} \\ \nu_{M_2} \end{bmatrix} \mid M_3 \sim N\left(\begin{bmatrix} \mu_\theta(M_3) \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma^2_\theta(M_3) & 0 & 0 \\ 0 & \sigma^2_{\nu_{M_1}}(M_3) & 0 \\ 0 & 0 & \sigma^2_{\nu_{M_2}}(M_3) \end{bmatrix} \right).$$

In order to pin down the location and scale of θ, I have normalized $\mu_{M_1} = 0$ and $\lambda_{M_1} = 1$ without loss of generality.

The latent outcome model takes the following form:

$$\begin{aligned} E[Y(0) \mid \theta] &= \alpha_0 + \beta_0\theta \\ E[Y(1) \mid \theta] &= \alpha_1 + \beta_1\theta \end{aligned}$$

I assume that the potential outcomes $Y(0)$ and $Y(1)$ are conditionally independent of M given θ. Formally,

$$(Y(0), Y(1)) \perp\!\!\!\perp M \mid \theta.$$

This means that the noisy measures of θ are related to the potential outcomes only through θ. Consequently, they can be used as instruments to identify the relationship between the potential outcomes and θ.

Given the simple parametric specification, the identification of $f_{\theta,M}$ depends on one's ability to identify the unknown parameters μ_{M_2} and λ_{M_2}, μ_θ, σ^2_θ, $\sigma^2_{\nu_{M_1}}$, and $\sigma^2_{\nu_{M_2}}$. These can be obtained from the moments

of the joint distribution of M_1, M_2, and M_3 by noticing that

$$\begin{aligned}
E[M_1 \mid M_3] &= \mu_\theta(M_3) \\
E[M_2] &= \mu_{M_2} + \lambda_{M_2} E[\mu_\theta(M_3)] \\
Var[M_1 \mid M_3] &= \sigma_\theta^2(M_3) + \sigma_{\nu_{M_1}}^2(M_3) \\
Var[M_2 \mid M_3] &= \lambda_{M_2}^2 \sigma_\theta^2(M_3) + \sigma_{\nu_{M_2}}^2(M_3) \\
Cov[M_1, M_2 \mid M_3] &= \lambda_{M_2} \sigma_\theta^2(M_3) \\
Cov[M_1, M_3] &= Cov[\theta, M_3] \\
Cov[M_2, M_3] &= \lambda_{M_2} Cov[\theta, M_3].
\end{aligned}$$

As long as $\lambda_{M_2}, Cov(\theta, M_3) \neq 0$, as was assumed above, the unknown parameters are given by

$$\begin{aligned}
\mu_\theta(M_3) &= E[M_1 \mid M_3] \\
\lambda_{M_2} &= \frac{Cov[M_2, M_3]}{Cov[M_1, M_3]} \\
\mu_{M_2} &= E[M_2] - \lambda_{M_2} E[\mu_\theta(M_3)] \\
\sigma_\theta^2(M_3) &= \frac{Cov[M_1, M_2 \mid M_3]}{\lambda_{M_k}} \\
\sigma_{\nu_{M_1}}^2(M_3) &= Var[M_1 \mid M_3] - \sigma_\theta^2(M_3) \\
\sigma_{\nu_{M_2}}^2(M_3) &= Var[M_2 \mid M_3] - \lambda_{M_2}^2 \sigma_\theta^2(M_3).
\end{aligned}$$

These parameters fully characterize the conditional joint distribution of θ, M_1, and M_2 given M_3. The joint distribution $f_{\theta,M}$ is then given by

$$f_{\theta,M}(\theta, m) = f_{\theta,M_1,M_2|M_3}(\theta, m_1, m_2 \mid m_3) f_{M_3}(m_3).$$

Let us now turn to the identification of $E[Y(1) - Y(0) \mid \theta]$. Given the simple parametric specifications for $E[Y(0) \mid \theta]$ and $E[Y(1) \mid \theta]$, the identification of $E[Y(1) - Y(0) \mid \theta]$ depends on one's ability to identify the unknown parameters α_0, β_0, α_1, and β_1. These can be obtained from the moments of the conditional distribution of Y given M and D and the conditional distribution of θ given M and D by noticing that

$$\begin{aligned}
E[Y \mid M = m^0, D = 0] &= E[Y(0) \mid M = m^0, D = 0] \\
&= E\{E[Y(0) \mid \theta] \mid\mid M = m^0, D = 0\} \\
&= \alpha_0 + \beta_0 E[\theta \mid M = m^0, D = 0] \\
E[Y \mid M = m^1, D = 1] &= E[Y(1) \mid M = m^1, D = 1] \\
&= E\{E[Y(1) \mid \theta] \mid\mid M = m^1, D = 1\} \\
&= \alpha_1 + \beta_1 E[\theta \mid M = m^1, D = 1]
\end{aligned}$$

for all $m^0 \in \mathcal{M}^0$ and $m^1 \in \mathcal{M}^1$.

The unknown parameters are given by

$$\begin{bmatrix} \alpha_0 \\ \beta_0 \end{bmatrix} = \begin{bmatrix} 1 & E\left[\theta \mid M = m^{0,1}, D = 0\right] \\ 1 & E\left[\theta \mid M = m^{0,2}, D = 0\right] \end{bmatrix}^{-1} \begin{bmatrix} E\left[Y \mid M = m^{0,1}, D = 0\right] \\ E\left[Y \mid M = m^{0,2}, D = 0\right] \end{bmatrix}$$
$$\begin{bmatrix} \alpha_1 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} 1 & E\left[\theta \mid M = m^{1,1}, D = 1\right] \\ 1 & E\left[\theta \mid M = m^{1,2}, D = 1\right] \end{bmatrix}^{-1} \begin{bmatrix} E\left[Y \mid M = m^{1,1}, D = 1\right] \\ E\left[Y \mid M = m^{1,2}, D = 1\right] \end{bmatrix}$$

where $m^{0,1}, m^{0,2} \in \mathcal{M}^0$ and $m^{1,1}, m^{1,2} \in \mathcal{M}^1$.These parameters fully characterize $E\left[Y\left(0\right) \mid \theta\right]$, $E\left[Y\left(1\right) \mid \theta\right]$, and consequently $E\left[Y\left(1\right) - Y\left(0\right) \mid \theta\right]$. The above result requires that the matrices

$$\begin{bmatrix} 1 & E\left[\theta \mid M = m^{0,1}, D = 0\right] \\ 1 & E\left[\theta \mid M = m^{0,2}, D = 0\right] \end{bmatrix}$$
$$\begin{bmatrix} 1 & E\left[\theta \mid M = m^{1,1}, D = 1\right] \\ 1 & E\left[\theta \mid M = m^{1,2}, D = 1\right] \end{bmatrix}$$

are full rank which is implied by the assumptions on the measurement model.

Finally, the conditional expectation functions of $Y\left(0\right)$ and $Y\left(1\right)$ given $R = r$, $E\left[Y\left(0\right) \mid R = r\right]$ and $E\left[Y\left(1\right) \mid R = r\right]$ as well as the conditional Average Treatment Effect given $R = r$, $E\left[Y\left(1\right) - Y\left(0\right) \mid R = r\right]$, are given by

$$\begin{aligned} E\left[Y\left(0\right) \mid R = r\right] &= E\left\{E\left[Y\left(0\right) \mid \theta\right] \mid R = r\right\} \\ &= \alpha_0 + \beta_0 E\left[\theta \mid R = r\right] \\ E\left[Y\left(1\right) \mid R = r\right] &= E\left\{E\left[Y\left(1\right) \mid \theta\right] \mid R = r\right\} \\ &= \alpha_1 + \beta_1 E\left[\theta \mid R = r\right] \\ E\left[Y\left(1\right) - Y\left(0\right) \mid R = r\right] &= \left(\alpha_1 - \alpha_0\right) + \left(\beta_1 - \beta_0\right) E\left[\theta \mid R = r\right] \end{aligned}$$

for all $r \in \mathcal{R}$.

In this example I have imposed strong parametric assumptions to illustrate the identification of the latent factor model. However, these assumptions are not necessary for identification. In the following sections I relax the distributional and functional form assumptions on the measurement model as well as the functional form assumptions on the latent outcome model.

1.2.3 Identification of the Latent Factor Distribution

1.2.3.1 Linear Measurement Model

I continue to assume a linear measurement model for M_1 and M_2 but leave the measurement model for M_3 flexible. The measurement model takes the following form:

$$M_1 = \theta + \nu_{M_1} \tag{1.1}$$

$$M_2 = \mu_{M_2} + \lambda_{M_2}\theta + \nu_{M_2} \tag{1.2}$$

$$M_3 = g_{M_3}(\theta, \nu_{M_3})$$

where $\lambda_{M_2} \neq 0$, $E[\nu_{M_1} \mid \theta] = E[\nu_{M_2} \mid \theta] = 0$. Assumption C lists conditions under which one can obtain nonparametric identification of $f_{\theta,M}$ using standard results from the literature on latent factor models as well as an extension to Kotlarski's Lemma (Kotlarski, 1967; Prakasa Rao, 1992) by Evdokimov and White (2012).

Assumption C.

1. The relationship between M_1, M_2, and θ is as given in equations (1.1) and (1.2).

2. θ, ν_{M_1}, and ν_{M_2} are jointly independent conditional on M_3.

3. $Cov[\theta, M_3] \neq 0$ and $E[|\theta| \mid M_3] < \infty$, $E[\nu_{M_1} \mid M_3] = E[\nu_{M_2} \mid M_3] = 0$ a.s.

4. One of the following conditions holds:

(a) The real zeros of the conditional characteristic function of ν_{M_1} given M_3 and its derivative are disjoint, and the conditional characteristic function of ν_{M_2} given M_3 has only isolated real zeros.

(b) The conditional characteristic function of ν_{M_1} given M_3 is analytic.

Assumption C.1 imposes linearity on the measurement models for M_1 and M_2 as discussed above. To pin down the location and scale of θ, I again use the normalization $\mu_{M_1} = 0$ and $\lambda_{M_2} = 1$. In addition, I assume that $\lambda_{M_2} \neq 0$ to guarantee that M_2 contains information about θ. Assumption C.2 restricts θ, ν_{M_1}, and ν_{M_3} to be jointly independent conditional on M_3. An important implication of this is that there cannot be heteroscedasticity in ν_{M_1} and ν_{M_2} with respect to θ. Assumption C.3 requires that M_3 is correlated with θ and that both ν_{M_1} and ν_{M_3} are mean independent of M_3. In addition, I assume that the conditional mean of θ given M_3 exists, thus ruling out distributions with particularly fat tails (e.g. the Cauchy distribution).

Lastly, Assumption C.4 imposes restrictions on the conditional characteristic functions of ν_{M_1} and ν_{M_2} given M_3. This assumption is best understood by considering first the original Kotlarski's Lemma (Kotlarski, 1967; Prakasa Rao, 1992). This lemma requires that the conditional characteristic functions of θ, ν_{M_1}, and ν_{M_2} given M_3 do not have any real zeros (such characteristic functions are typically called nonvanishing).[11] This is a common assumption in the measurement error literature (Chen, Hong, and Nekipelov, 2011). It is

[11] A nonvanishing characteristic function is closely related to (bounded) completeness of a location family. Therefore, it can be seen as requiring that the distribution varies sufficiently as a function of the location parameter. See the related discussion and references in Section 1.2.3.2.

satisfied by most standard distributions, such as the normal, log-normal, Cauchy, Gamma, Laplace, χ^2 and Student's t-distribution. However, it is violated by, for instance, uniform, triangular, and truncated normal distributions as well as by many discrete distributions.

Assumption C.4 uses recent work by Evdokimov and White (2012) to relax the assumptions of Kotlarski's Lemma. Condition (a) allows for real zeros in the conditional characteristic functions of ν_{M_1} and ν_{M_2} given M_3. This substantially expands the class of distributions that allow for the identification of $f_{\theta,M}$. Condition (b) requires the conditional characteristic function of ν_{M_1} given M_3 to be analytic while imposing no restrictions on the conditional characteristic function of ν_{M_2} given M_3. Analyticity is a property satisfied, for instance, by distributions with exponentially bounded tails, such as the normal and Gamma distribution as well as distributions with bounded support. Importantly, neither one of these conditions imposes restrictions on the conditional characteristic function of θ given M_3.

Theorem 1 states the identification result. Given the normalizations imposed on the linear measurement model, the covariance and mean independence assumptions imply that λ_{M_2} is given by the ratio of the covariance between M_2 and M_3 and the covariance between M_1 and M_3. This amounts to using M_3 as an instrument for M_1 in a regression of M_2 on M_1 as M_1 is an imperfect measure of θ. The means of M_1 and M_2 can then be used to obtain μ_{M_2}. Suppose for a moment that the conditional distributions of θ, ν_{M_1}, and ν_{M_2} given M_3 were known up to a finite number of parameters. In this case one could, subject to the relevant full rank condition, identify the distributional parameters from a finite number of moment conditions. The rest of Theorem 1 uses the assumptions on the conditional characteristic functions to generalize this strategy to a infinite-dimensional problem. Under these assumptions, the conditional distributions of ν_{M_1} and ν_{M_2} given M_3, $f_{\nu_{M_1}|M_3}$ and $f_{\nu_{M_2}|M_3}$, as well as the conditional distribution of θ given M_3, $f_{\theta|M_3}$, can be uniquely determined from the conditional distribution of M_1 and M_2 given M_3, $f_{M_1,M_2|M_3}$. Together with the marginal distribution of M_3, f_{M_3}, this allows one to then construct $f_{\theta,M}$.

Theorem 1. *Suppose Assumption C holds. Then*

$$\begin{aligned}\mu_{M_2} &= E[M_2] - \lambda_{M_2} E[M_1] \\ \lambda_{M_2} &= \frac{Cov[M_2, M_3]}{Cov[M_1, M_3]}.\end{aligned}$$

In addition, the equation

$$\begin{aligned}&f_{M_1,M_2|M_3}(m_1, m_2 \mid m_3) \\ &= \int_{\Theta} f_{\nu_{M_1}|M_3}(m_1 - \theta \mid m_3)\, f_{\nu_{M_2}|M_3}(m_2 - \mu_{M_2} - \lambda_{M_2}\theta \mid m_3)\, f_{\theta|M_3}(\theta \mid m_3)\, d\theta\end{aligned}$$

for all $m_1 \in \mathcal{M}_1$, $m_2 \in \mathcal{M}_2$, and $m_3 \in \mathcal{M}_3$ admits unique solutions for $f_{\nu_{M_1}|M_3}$, $f_{\nu_{M_2}|M_3}$, and $f_{\theta|M_3}$. Consequently, the joint distribution $f_{\theta,M}$ is identified.

1.2.3.2 Nonlinear Measurement Model

The linear measurement model considered in Section 1.2.3.1 provides a natural starting point for the discussion regarding the identification of $f_{\theta,M}$. However, this model imposes important, and potentially unsatisfactory, restrictions on the relationship between the latent factor θ and the measures M_1 and M_2. First, both M_1 and M_2 are assumed to depend linearly on θ. Second, θ, ν_{M_1}, and ν_{M_2} are assumed to be jointly independent conditional on M_3. This rules out, for instance, heteroskedasticity in ν_{M_1} and ν_{M_2} with respect to θ. In this section I discuss an alternative set of identifying assumptions that address these concerns. For this purpose I use results by Hu and Schennach (2008) who study nonparametric identification and estimation in the presence of nonclassical measurement error.

I return to the general measurement model that took the form

$$\begin{aligned} M_1 &= g_{M_1}(\theta, \nu_{M_1}) \\ M_2 &= g_{M_2}(\theta, \nu_{M_2}) \\ M_3 &= g_{M_3}(\theta, \nu_{M_3}). \end{aligned}$$

Assumption D lists the conditions under which the joint distribution $f_{\theta,M}$ is nonparametrically identified in this setting (Hu and Schennach, 2008; Cunha, Heckman, and Schennach, 2010).

Assumption D.

1. $f_{\theta,M}(\theta, m)$ *is bounded with respect to the product measure of the Lebesgue measure on* $\Theta \times \mathcal{M}_1 \times \mathcal{M}_2$ *and some dominating measure* μ *on* $\mathcal{M}_3$*. All the corresponding marginal and conditional densities are also bounded.*

2. M_1*,* M_2*, and* M_3 *are jointly independent conditional on* θ*.*

3. For all $\theta', \theta'' \in \Theta$*,* $f_{M_3|\theta}\left(m_3 \mid \theta'\right)$ *and* $f_{M_3|\theta}\left(m_3 \mid \theta''\right)$ *differ over a set of strictly positive probability whenever* $\theta' \neq \theta''$*.*

4. There exists a known functional H *such that* $H\left[f_{M_1|\theta}(\cdot \mid \theta)\right] = \theta$ *for all* $\theta \in \Theta$*.*

5. $f_{\theta|M_1}(\theta \mid m_1)$ *and* $f_{M_1|M_2}(m_1 \mid m_2)$ *form (boundedly) complete families of distributions indexed by* $m_1 \in \mathcal{M}_1$ *and* $m_2 \in \mathcal{M}_2$*.*

Assumption D.1 requires θ, M_1, and M_2 to be continuos but allows M_3 to be either continuous or discrete. Furthermore, it restricts the joint, marginal and conditional densities of θ, M_1, M_2, and M_3 to be bounded. The support of the joint distribution $f_{\theta,M}$, on the other hand, is allowed to be either rectangular or triangular. Assumption D.2 restricts ν_{M_1}, ν_{M_2}, and ν_{M_3} to be jointly independent conditional on θ while allowing for arbitrary dependence between θ and these disturbances. Importanly, this assumption allows for both heteroscedasticity and correlation in the measurement errors in M_1, M_2, and M_3. Assumption D.3 requires the conditional distribution of M_3 given θ to vary sufficiently as a function of θ. This assumption can be satisfied, for instance, by assuming strict monotonicity of the conditional expectation of M_3 given

θ. More generally this assumption can be satisfied if there is heteroscedasticity in M_3 with respect to θ. Assumption D.4 imposes a normalization on the conditional distribution of M_1 given θ in order to pin down the location and scale of θ. This normalization can be achieved by, for instance, requiring the conditional mean, mode or median of M_1 to be equal to θ.

Lastly, Assumption D.5 requires that the conditional distributions $f_{\theta|M_1}$ and $f_{M_1|M_2}$ are either complete or boundedly complete.[12] The concept of completeness, originally introduced in statistics by Lehmann and Scheffe (1950, 1955), arises regularly in econometrics, for instance, as a necessary condition for the identification of (semi-)nonparametric instrumental variable models (Newey and Powell, 2003; Blundell and Powell, 2003; Chernozhukov and Hansen, 2005; Blundell, Chen, and Kristensen, 2007; Chernozhukov, Imbens, and Newey, 2007).[13] It can be seen as an infinite-dimensional generalization of the full rank condition that is central in the identification of various parametric models based on Generalized Method of Moments (Hansen, 1982).[14] Intuitively, the completeness condition requires that $f_{\theta|M_1}$ and $f_{M_1|M_2}$ vary sufficiently as functions of M_1 and M_2. One way to see this is to consider the assumption of L^2-completeness which lies in between completeness and bounded completeness in terms of its restrictiveness.[15] It can be shown that the conditional distribution of X given Z, where X and Z denote generic random variables, is L^2-complete if and only if every nondegenerate square-integrable function of X is correlated with some square-integrable function of Z (Severini and Tripathi, 2006; Andrews, 2011).

An unsatisfactory feature of completeness assumptions is that, unlike the full rank condition in finite-dimensional models, these assumptions are generally untestable (Canay, Santos, and Shaikh, forthcoming). However, there has been some work on providing sufficient conditions for various forms of completeness of certain classes of distributions, such as the location, scale and exponential families, in both statistics (Ghosh and Singh, 1966; Isenbeck and Ruschendorf, 1992; Mattner, 1992; Lehmann and Romano, 2005) and econometrics (D'Haultfoeuille, 2011; Hu and Shiu, 2012).[16] In addition, some papers in the literature have focused on characterizing classes of distributions that fail the completeness assumption but satisfy the weaker bounded completeness assumption (Hoeffding, 1977; Bar-Lev and Plachky, 1989; Mattner, 1993). Finally, Andrews (2011) and Chen, Chernozhukov, Lee, and Newey (2013) have provided genericity results that imply that L^2-completeness holds almost surely for large classes of nonparametric distributions.[17]

[12] Let X and Z denote generic random variables with supports $\mathcal{X}$ and $\mathcal{Z}$. $f_{X|Z}(x \mid z)$ is said to form a (boundedly) complete family of distributions indexed by $z \in \mathcal{Z}$ if for all measurable (bounded) real functions h such that $E[h(X)] < \infty$, $E[h(X) \mid Z] = 0$ a.s. implies $h(X) = 0$ a.s. (Lehmann and Romano, 2005).

[13] Completeness is sometimes stated in the literature on (semi-)nonparametric instrumental variable models in terms of injectivity of the conditional expectation operator (Hall and Horowitz, 2005; Darolles, Fan, Florens, and Renault, 2011; Horowitz, 2011). Some authors also refer to completeness as strong identification (Florens, Mouchart, and Rolin, 1990).

[14] Notice that if the generic random variables X and Z are both discrete with finite supports $\mathcal{X} = \{x_1, \ldots, x_K\}$ and $\mathcal{Z} = \{z_1, \ldots, z_L\}$, the completeness assumption becomes the full rank condition $P[rank(Q) = K] = 1$ where $Q_{kl} = P[X = x_k \mid Z = z_l]$ (Newey and Powell, 2003).

[15] The definition of L^2-completeness is analogous to the definition given above for (bounded) completeness with the exception that the condition needs to hold for all measurable square-integrable real functions h (Andrews, 2011). Thus, L^2-completeness lies in between completeness and bounded completeness in the sense that completeness implies L^2-completeness which in turn implies bounded completeness.

[16] For instance, the assumption of a nonvanishing characteristic function discussed in Section 1.2.3.1 is a necessary condition for completeness and a necessary and sufficient condition for bounded completeness of a location family (Ghosh and Singh, 1966; Isenbeck and Ruschendorf, 1992; Mattner, 1992).

[17] See also Santos (2012) for a related discussion on the uniform closeness of complete and incomplete distributions.

Theorem 2 states the identification result by Hu and Schennach (2008). Given the conditional joint independence assumption, one can write down the integral equation given in the theorem relating the observed conditional joint distribution of M_1 and M_3 given M_3, $f_{M_1,M_3|M_2}$, to the unobserved distributions $f_{M_1|\theta}$, $f_{M_3|\theta}$, and $f_{\theta|M_2}$. Futhermore, this relationship can be expressed in terms of linear integral operators that makes the problem analogous to matrix diagonalization in linear algebra. Using invertibility of some of the operators, provided by the (bounded) completeness assumption, one can obtain an eigenvalue-eigenfunction decomposition of an integral operator that only depends on the observed f_M. Given the additional assumptions, this decomposition is unique, and the unknown densities $f_{M_1|\theta}$, $f_{M_3|\theta}$, and $f_{\theta|M_2}$ are given by the eigenfunctions and eigenvalues of this decomposition. This allows one to then construct $f_{\theta,M}$.

Theorem 2. *Suppose Assumption D holds. Then the equation*

$$f_{M_1,M_3|M_2}(m_1,m_3 \mid m_2) = \int_\Theta f_{M_1|\theta}(m_1 \mid \theta) f_{M_3|\theta}(m_3 \mid \theta) f_{\theta|M_2}(\theta \mid m_2)\, d\theta$$

for all $m_1 \in \mathcal{M}_1$, $m_2 \in \mathcal{M}_2$ and $m_3 \in \mathcal{M}_3$ admits unique solutions for $f_{M_1|\theta}$, $f_{M_3|\theta}$, and $f_{\theta|M_2}$. Consequently, the joint distribution $f_{\theta,M}$ is identified.

1.2.4 Identification of the Latent Conditional Average Treatment Effect

Having identified the joint distribution of θ and M, $f_{\theta,M}$, and the conditional distribution of θ given R, $f_{\theta|R}$, the only missing piece in the identification of $E[Y(1) - Y(0) \mid R]$ is the latent conditional Average Treatment Effect $E[Y(1) - Y(0) \mid \theta]$. The identification of this is based on the identification of the latent conditional expectation functions $E[Y(0) \mid \theta]$ and $E[Y(1) \mid \theta]$. These functions can be identified by relating the variation in the conditional expectation of Y given M and D, $E[Y \mid M, D]$, to the variation in the conditional distribution of θ given M and D, $f_{\theta|M,D}$ to the left ($D = 0$) and right ($D = 1$) of the cutoff. This problem is analogous to the identification of separable (semi-)nonparametric instrumental variable models (Newey and Powell, 2003; Darolles, Fan, Florens, and Renault, 2011). Assumption E lists the conditions under which $E[Y(0) \mid \theta]$ and $E[Y(1) \mid \theta]$ are nonparametrically identified for all $\theta \in \Theta$.

Assumption E.

1. $(Y(0), Y(1)) \perp\!\!\!\perp M \mid \theta$.

2. $0 < P[D = 1 \mid \theta] < 1$ *a.s.*

3. $f_{\theta|M,D}(\theta \mid m^0, 0)$ *and* $f_{\theta|M,D}(\theta \mid m^1, 1)$ *form (boundedly) complete families of distributions indexed by* $m^0 \in \mathcal{M}^0$ *and* $m^1 \in \mathcal{M}^1$.

Assumption E.1 requires that the potential outcomes $Y(0)$ and $Y(1)$ are conditionally independent of the measures M given the latent factor θ. In other words, the measurement errors in M are not allowed to affect $Y(0)$ and $Y(1)$.[18] Assumption E.2 states a common support condition that guarantees that the

[18]This condition is often referred to as nondifferential measurement error (Bound, Brown, and Mathiowetz, 2001; Carroll, Ruppert, Stefanski, and Crainiceanu, 2006).

conditional supports of θ to the left and right of the cutoff c coincide. This means that the subset of M entering R must be sufficiently noisy measures of θ so that for all $\theta \in \Theta$ the realized value of R can lie on both sides of the cutoff with strictly positive probability.

Finally, Assumption E.3 imposes a similar (bounded) completeness condition as in Assumption D for the identification of the nonlinear measurement model. Here the vector M is used as an instrument for θ, and the (bounded) completeness conditions can be thought of as an infinite-dimensional first stage condition. A sufficient condition for this assumption to be implied by the (bounded) completeness conditions in Assumption D is that M_1 does not enter R, and that there exists some $\left(m_2^d, m_3^d\right) \in \mathcal{M}_2^d \times \mathcal{M}_3^d$ such that $f_{\theta,M_2,M_3|D}\left(\theta, m_2^d, m_3^d \mid d\right) > 0$ for all $\theta \in \Theta$, $d = 0, 1$.

Theorem 3 states the identification result. The conditional independence assumption allows one to write down the integral equations given in the theorem. Under the (bounded) completeness assumption, $E[Y(0) \mid \theta]$ and $E[Y(1) \mid \theta]$ are unique solutions to these integral equations. Finally, the common support assumption ensures that both $E[Y(0) \mid \theta]$ and $E[Y(1) \mid \theta]$ are determined for all $\theta \in \Theta$.

Theorem 3. *Suppose Assumption E holds. Then the equations*

$$\begin{aligned} E\left[Y \mid M = m^0, D = 0\right] &= E\left\{E\left[Y(0) \mid \theta\right] \mid M = m^0, D = 0\right\} \\ E\left[Y \mid M = m^1, D = 1\right] &= E\left\{E\left[Y(1) \mid \theta\right] \mid M = m^1, D = 1\right\} \end{aligned}$$

for all $m^0 \in M^0$ and $m^1 \in \mathcal{M}^1$ admit unique solutions for (bounded) $E[Y(0) \mid \theta]$ and $E[Y(1) \mid \theta]$ for all $\theta \in \Theta$. Consequently, $E[Y(1) - Y(0) \mid \theta]$ is identified.

1.3 Extensions

1.3.1 Extrapolation of Local Average Treatment Effect in Fuzzy RD

In fuzzy RD the treatment is only partly determined by whether the running variable falls above or below the cutoff c: some individuals assigned to the treatment may end up not receiving the treatment while some individuals not assigned to the treatment may end up receiving the treatment. Thus, in fuzzy RD the probability of receiving the treatment jumps when the running variable R crosses the cutoff c but by less than 1:

$$\lim_{\delta \downarrow 0} P[D = 1 \mid R = r + \delta] > \lim_{\delta \downarrow 0} P[D = 1 \mid R = r - \delta].$$

Let Z denote the treatment assignment that is a deterministic function of the running variable:

$$Z = 1(R \geq c)$$

Each individual is associated with two potential treatment status: $D(0)$ is the treatment status of an individual if she is not assigned to the treatment ($Z = 0$), and $D(1)$ is the treatment status of an individual if she is assigned to the treatment ($Z = 1$). Using this notation, the observed outcome and the observed treatment status can be written as

$$\begin{aligned} Y &= Y(0) + (Y(1) - Y(0))\, D \\ D &= D(0) + (D(1) - D(0))\, Z. \end{aligned}$$

It is possible to categorize individuals into four mutually exclusive groups according to their compliance with treatment assignment (Imbens and Angrist, 1994; Angrist, Imbens, and Rubin, 1996): (1) individuals who receive the treatment whether or not they are assigned to the treatment are called always-takers ($D(0) = D(1) = 1$), (2) individuals who do not receive the treatment whether or not they are assigned to the treatment are called never-takers ($D(0) = D(1) = 0$), (3) individuals who receive the treatment if they are assigned to the treatment and do not receive the treatment if they are not assigned to the treatment are called compliers ($D(0) = 0, D(1) = 1$), and (4) individuals who receive the treatment if they are not assigned to the treatment and do not receive treatment if they are asssigned to the treatment are called defiers ($D(0) = 1, D(1) = 0$).

I rule out defiers by assuming that being assigned to the treatment can only make an individual more likely to receive the treatment. This corresponds to the monotonicity assumption in the instrumental variables literature (Imbens and Angrist, 1994; Angrist, Imbens, and Rubin, 1996). Once defiers have been ruled out, fuzzy RD allows one to nonparametrically identify the Local Average Treatment Effect (LATE) for the compliers at the cutoff, $E[Y(1) - Y(0) \mid D(1) > D(0), R = c]$. This is the group of individuals whose treatment status changes at the cutoff as they become eligible to the treatment. Since the treatment status of never-takers and always-takers is independent of treatment assignment, fuzzy RD contains no information about the Average Treatment Effect for these two groups.

Assumption F lists conditions under which the Local Average Treatment Effect is nonparametrically identified. Assumption F.1 restricts the marginal density of R, f_R, to be strictly positive in a neighborhood around the cutoff. Assumption F.2 imposes the monotonicity assumption stating that crossing the cutoff can only make an individual more likely to receive the treatment. In addition, it requires that this relationship is strict for at least some individuals at the cutoff, ensuring that there is a first stage. Assumption F.3 requires the conditional expectations of the potential treatment status to be continuous in R at the cutoff. Assumption F.4 imposes continuity on the conditional cumulative distribution functions of the potential outcomes for the compliers.[19] Finally, Assumption F.5 requires that the conditional expectations of $Y(0)$ and $Y(1)$ exist at the cutoff.

Assumption F.

[19] Continuity of the conditional expectation functions of the potential outcomes for compliers is enough for Lemma 3. However, continuity of the conditional cumulative distribution functions allows one to also identify distributional treatment effects for the compliers (Frandsen, Frolich, and Melly, 2012).

1. $f_R(r) > 0$ *in a neighborhood around* c.
2. $\lim_{\delta \downarrow 0} P[D(1) \geq D(0) \mid R = r \pm \delta] = 1$ *and* $\lim_{\delta \downarrow 0} P[D(1) > D(0) \mid R = r \pm \delta] > 0$
3. $E[D(0) \mid R = r]$ *and* $E[D(1) \mid R = r]$ *are continuous in* r *at* c.
4. $F_{Y(0)|D(0),D(1),R}(y \mid 0, 1, r)$ *and* $F_{Y(1)|D(0),D(1),R}(y \mid 0, 1, r)$ *are continuous in* r *at* c *for all* $y \in \mathcal{Y}$.
5. $E[Y(0) \mid D(1) > D(0), R = c], E[Y(0) \mid D(1) > D(0), R = c] < \infty$.

The identification result is given in Lemma 3 (Hahn, Todd, and van der Klaauw, 2001). Under Assumption F, the Local Average Treatment Effect can be obtained as the ratio of the difference in the limits of the conditional expectation of Y given $R = r$, $E[Y \mid R = r]$, as r approaches c from right and left, and the difference in the limits of the conditional expectation of D given $R = r$, $E[D \mid R = r]$, as r approaches c from right and left. In other words, any discontinuity observed at the cutoff in the conditional expectation of Y given R is accredited to the treatment through the corresponding discontinuity at the cutoff in the probability of receiving the treatment.

Lemma 3. *(Hahn, Todd, and van der Klaauw, 2001) Suppose Assumption F holds. Then*

$$E[Y(1) - Y(0) \mid D(1) > D(0), R = c] = \lim_{\delta \downarrow 0} \frac{E[Y \mid R = c + \delta] - E[Y \mid R = c - \delta]}{E[D \mid R = c + \delta] - E[D \mid R = c - \delta]}.$$

Assumption G lists conditions under which the Local Average Treatment Effect for compliers at any point r in the running variable distribution, $E[Y(1) - Y(0) \mid D(1) > D(0), R = r]$, is nonparametrically identified in the latent factor framework. Assumption G.1 requires that the potential outcomes $Y(0)$ and $Y(1)$ and the potential treatment status $D(0)$ and $D(1)$ are jointly independent of R conditional on the latent factor θ. Assumption G.2 imposes the monotonicity assumption for all $\theta \in \Theta$. Assumption G.3 imposes this relationship to be strict at least for some $\theta \in \Theta$.

Assumption G.

1. $(Y(0), Y(1), D(0), D(1)) \perp\!\!\!\perp R \mid \theta$.
2. $P[D(1) \geq D(0) \mid \theta] = 1$ *a.s.*
3. $P[D(1) > D(0) \mid \theta] > 0$ *a.s.*

The identification result is stated in Lemma 4. Under Assumption G, the Local Average Treatment Effect for complier at $R = r$ is given by the ratio of the reduced from effect of treatment assignment on the outcome and the first stage effect of treatment assignment on treatment status at $R = r$.

Lemma 4. *Suppose Assumption G holds. Then*

$$E[Y(1) - Y(0) \mid D(1) > D(0), R = r] = \frac{E\{E[Y(D(1)) - Y(D(0)) \mid \theta] \mid R = r\}}{E\{E[D(1) - D(0) \mid \theta] \mid R = r\}}$$

for all $r \in \mathcal{R}$.

Assumption H lists the conditions under which the latent reduced form effect of treatment assignment on the outcome, $E[Y(D(1)) - Y(D(0)) \mid \theta]$, and the latent first stage effect of treatment assignment on the probability of receiving the treatment, $E[D(1) - D(0) \mid \theta]$, are nonparametrically identified from the conditional distribution of Y given M and Z, $f_{Y|M,Z}$, and the conditional distribution of D given M and Z, $f_{D|M,Z}$. Assumption H.1 requires that the potential outcomes $Y(0)$ and $Y(1)$ and the potential treatment status $D(0)$ and $D(1)$ are jointly independent of M given θ. In other words, the measurement errors in M are assumed to to be unrelated to $(Y(0), Y(1), D(0), D(1))$. Assumption H.2 repeats the common support assumption from Assumption E whereas Assumption H.3 is analogous to the (bounded) completeness condition in Assumption E.

Assumption H.

1. $(Y(0), Y(1), D(0), D(1)) \perp\!\!\!\perp M \mid \theta$.

2. $0 < P[D = 1 \mid \theta] < 1$ *a.s.*

3. $f_{\theta|M,Z}(\theta \mid m^0, 0)$ *and* $f_{\theta|M,Z}(\theta \mid m^1, 1)$ *form (boundedly) complete families of distributions indexed by* $m^0 \in \mathcal{M}^0$ *and* $m^1 \in \mathcal{M}^1$.

Theorem 4 states the identification result. Together with Lemma 4 this result can be used to nonparametrically identify the Local Average Treatment Effect for compliers at any point in the running variable distribution. The proof of Theorem 4 is analogous to the proof of Theorem 3.

Theorem 4. *Suppose Assumption H holds. Then the equations*

$$\begin{aligned}
E[Y \mid M = m^0, D = 0] &= E\{E[Y(D(0)) \mid \theta] \mid M = m^0, D = 0\} \\
E[Y \mid M = m^1, D = 1] &= E\{E[Y(D(1)) \mid \theta] \mid M = m^1, D = 1\} \\
E[D \mid M = m^0, D = 0] &= E\{E[D(0) \mid \theta] \mid M = m^0, D = 0\} \\
E[D \mid M = M^1, D = 1] &= E\{E[D(1) \mid \theta] \mid M = m^1, D = 1\}
\end{aligned}$$

for all $r_0 \in \mathcal{R}_0$ *and* $r_1 \in \mathcal{R}_1$ *admit unique solutions for (bounded)* $E[Y(D(0)) \mid \theta]$, $E[Y(D(1)) \mid \theta]$, $E[D(0) \mid \theta]$, *and* $E[D(1) \mid \theta]$ *for all* $\theta \in \Theta$.

1.3.2 Settings with Multiple Latent Factors

Section 1.2 focused on the identification of the measurement and latent outcome models in the presence of a one-dimensional latent factor θ. However, it is possible to generalize the identification results to a setting with a K-dimensional latent factor $\theta = (\theta_1, \ldots, \theta_K)$. Instead of three noisy measures required in the one-dimensional case, the K-dimensional case requires the availability of $2 \times K + 1$ noisy measures. To be more exact, this setting requires one to observe two noisy measures for each latent factor θ_k, $k = 1, \ldots, K$, as well as one measure that is related to all K latent factors.[20]

[20] It is possible to allow for this measure to be multidimensional by, for instance, containing an additional K measures for each latent factor θ_k, $k = 1, \ldots, K$.

Formally, I assume that the data contains $2 \times K + 1$ noisy measures given by

$$\begin{aligned} M_1^k &= g_{M_1^k}\left(\theta_k, \nu_{M_1^k}\right), k = 1, \ldots, K \\ M_2^k &= g_{M_2^k}\left(\theta_k, \nu_{M_2^k}\right), k = 1, \ldots, K \\ M_3 &= g_W\left(\theta_1, \ldots, \theta_K, \nu_{M_3}\right) \end{aligned}$$

where $g_{M_1^k}$, $g_{M_2^k}$, $k = 1, \ldots, K$, and g_{M_3} are unknown functions, and $\nu_{M_1^k}$, $\nu_{M_2^k}$, $k = 1, \ldots, K$, and ν_{M_3} are potentially multidimensional disturbances. I focus on a setting in which is R is a deterministic function of at least one but potentially many of the measures for each θ_k, $k = 1, \ldots, K$. However, it is possible to consider a more general setting that allows the relationship between R and M to be stochastic. Going back to the selective school example of Section 1.2.1, one might think of R as being the average score in two entrance exams in English and Math, M_1^1 and M_2^1, that are noisy measures of English and Math ability, θ_1 and θ_2. M_1^2, M_2^2, and M_3 might instead consist of pre-application baseline test scores in English and Math.

I require θ_k, M_1^k, and M_2^k, $k = 1, \ldots, K$, to be continuous but allow M_3 to be either continuous or discrete; even a binary M_3 suffices for identification. I denote the supports of θ_k, M_1^k, M_2^k,$k = 1, \ldots, K$, and M_3 by $\times_k$, $\mathcal{M}_1^k$, $\mathcal{M}_2^k$, $k = 1, \ldots, K$, and $\mathcal{M}_3$. In addition, I use $\Theta = \Theta_1 \times \cdots \times \Theta_K$ to denote the support of $\theta = (\theta_1, \ldots, \theta_K)$, $\mathcal{M}_1 = \mathcal{M}_1^1 \times \cdots \times \mathcal{M}_1^K$ and$\mathcal{M}_2 = \mathcal{M}_2^1 \times \cdots \times \mathcal{M}_2^K$ to denote the supports of $M_1 = \left(M_1^1, \ldots, M_1^K\right)$ and $M_2 = \left(M_2^1, \ldots, M_2^K\right)$, and $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \mathcal{M}_3$ to denote the support of $M = (M_1, M_2, M_3)$.

The introduction of multiple latent factors only affects Assumption C and Theorem 1 regarding the identification of $f_{\theta,M}$ in the linear measurement model. Assumption D and Theorem 2 regarding the identification of $f_{\theta,M}$ in the nonlinear measurement model apply instead directly to this setting as long as one interprets θ and M as defined above (Hu and Schennach, 2008). The same holds for Assumption E and Theorem 3 regarding the identification of the Average Treatment Effect away from the cutoff in sharp RD as well as for Assumption H and Theorem 4 regarding the identification of the Local Average Treatment Effect away from the cutoff in fuzzy RD.

Thus, I focus here on the identification of $f_{\theta,M}$ in the linear measurement model

$$M_1^k = \theta_k + \nu_{M_1^k}, k = 1, \ldots, K \tag{1.3}$$

$$M_2^k = \mu_{M_2^k} + \lambda_{M_2^k}\theta_k + \nu_{M_2^k}, k = 1, \ldots, K \tag{1.4}$$

$$M_3 = g_{M_3}\left(\theta_1, \ldots, \theta_K, \nu_{M_3}\right)$$

where $\lambda_{M_2^k} \neq 0$, $E\left[\nu_{M_1^k} \mid \theta_k\right] = E\left[\nu_{M_2^k} \mid \theta_k\right] = 0$, $k = 1, \ldots, K$. Assumption I lists modified conditions under which one can obtain nonparametric identification of $f_{\theta,M}$ in this setting.

Assumption I.

1. The relationship between M_1^k, M_2^k, and θ_k, $k = 1, \ldots, K$, are as given in Equations (1.3) and (1.4).

2. θ_k, ν_{R_k}, *and* ν_{B_k} *are jointly independent conditional on* M_3 *for all* $k = 1, \ldots, K$.

3. $Cov[\theta_k, M_3] \neq 0$, $E[|\theta_k| \mid M_3] < \infty$, $E\left[\nu_{M_1^k} \mid M_3\right] = E\left[\nu_{M_1^k} \mid M_3\right] = 0$ *for all* $k = 1, \ldots, K$.

4. One of the following conditions holds:

(a) The real zeros of the conditional characteristic functions of $\nu_{M_1^k}$, $k = 1, \ldots, K$, *given* M_3 *and their derivatives are disjoint, and the conditional characteristic functions of* $\nu_{M_2^k}$, $k = 1, \ldots, K$, *given* M_3 *have only isolated real zeros.*

(b) The conditional characteristic functions of $\nu_{M_1^k}$, $k = 1, \ldots, K$, *given* M_3 *are analytic.*

5. The components of $\nu_{M_1} = \left(\nu_{M_1^1}, \ldots, \nu_{M_1^K}\right)$ *and* $\nu_{M_2} = \left(\nu_{M_2^1}, \ldots, \nu_{M_2^K}\right)$ *are jointly independent conditional on* M_3.

6. The real zeros of the conditional joint characteristic functions of ν_{M_1} *and* ν_{M_2} *given* M_3 *are disjoint.*

Theorem 5 states the identification result. The proof of this theorem is similar to the proof of Theorem 1.

Theorem 5. *Suppose Assumption I holds. Then*

$$\begin{aligned} \mu_{M_2^k} &= E\left[M_2^k\right] - \lambda_{M_2^k} E\left[M_1^k\right], \, k = 1, \ldots, K \\ \lambda_{M_2^k} &= \frac{Cov\left[M_2^k, M_3\right]}{Cov\left[M_1^k, M_3\right]}, \, k = 1, \ldots, K. \end{aligned}$$

In addition, the equation

$$\begin{aligned} & f_{M_1, M_1 | M_3}(m_1, m_2 \mid m_3) \\ &= \int_\Theta \left[\prod_{k=1}^K f_{\nu_{M_1^k} | M_3}\left(m_1^k - \theta_k \mid m_3\right) f_{\nu_{M_2^k} | M_3}\left(m_2^k - \mu_{M_2^k} - \lambda_{M_2^k} \theta_k \mid m_3\right) \right] \\ &\quad \times f_{\theta_k | M_3}(\theta_k \mid m_3) \, d\theta \end{aligned}$$

for all $m_1 \in \mathcal{M}_1$, $m_2 \in \mathcal{M}_2$, $m_3 \in \mathcal{M}_3$, *admits unique solutions for* $f_{\nu_{M_1^k} | M_3}$, $f_{\nu_{M_2^k} | M_3}$, $k = 1, \ldots, K$, *and* $f_{\theta | M_3}$. *Consequently, the joint distribution* $f_{\theta, M}$ *is identified.*

1.4 Boston Exam Schools

I use the latent factor-based approach to RD extrapolation developed in the previous two sections to study the causal effects of attending selective public schools, known as exam schools, in Boston for the full population of applicants. This section describes the empirical setting and data as well as the identification and estimation of the latent factor model in the empirical setting.

1.4.1 Setting

Boston Public Schools (BPS) includes three exam schools that span grades 7-12: Boston Latin School, Boston Latin Academy, and John D. O'Bryant High School of Mathematics and Science. Latin School, founded in 1635, is the oldest and most selective out of the three exam schools, and it is also the first public school and oldest still existing school in the United States. Latin School enrolls about 2,400 student. Latin Academy, founded in 1877, is the second oldest and second most selective exam school in Boston. It enrolls about 1,700 students. O'Bryant, founded in 1893, is the youngest and least selective out of the three exam schools. It enrolls about 1,200 students.

The exam schools differ considerably from traditional Boston public schools in terms of student performance. In the U.S. News & World Report high school ranking in 2013, for instance, Latin School, Latin Academy and O'Bryant formed the three best high schools in BPS, and ranked as 2nd, 20th, and 15th in Massachusetts. Furthermore, in 2012, the exam schools were among the four best schools in BPS in terms of the share of students scoring at a Proficient or Advanced level in the Massacusess Comprehensive Assesment System (MCAS) tests in English, Math, and Science.[21] Similarly, the exam schools formed the three best schools in BPS in 2012 in terms of both average SAT scores and 4-year graduation rates.[22]

The fact that exam school students, on average, outperform other BPS students in terms of MCAS/SAT scores and graduation rates is not surprising given the considerable differences in student composition between the exam schools and traditional Boston public schools. For instance, the exam schools enroll a considerably higher share of white and Asian students as well as a higher share of female students than BPS as a whole. Limited English proficiency, special education, and low income rates are also negligible among exam school students when compared to other Boston public schools.

In addition to student composition, the exam schools differ from traditional Boston public schools along several other dimensions. The exam schools have a higher share of teachers who are licensed to teach in the area in which they teach, as well as a higher share of core academic classes that are taught by teachers who hold a valid Massachusetts license and have demonstrated subject matter competency in the areas they teach. Exam school teachers are also older than teachers at other Boston public schools. The student/teacher ratio is much higher at the exam schools, but this is to a large extent explained by the small number of students requiring special education.

There are also considerable differences between the exam schools and traditional Boston public schools in terms of their curricula. The curriculum at both Latin Academy and Latin School emphasizes the classics, and students at these schools take mandatory Latin classes, whereas the curriculum at O'Bryant focuses on Math and Science. Moreover, the exam schools offer a rich array of college prepatory classes and extracurricular activities, and enjoy to a varying extent additional funding that come from alumni contributions.

[21] MCAS is a state-mandated series of achievement tests introduced for the purposes of No Child Left Behind.

[22] All of the exam schools are also among the few schools in BPS that have won the Blue Ribbon Award from the US Department of Education. In addition, Latin School was listed as one the top 20 high schools in the US by U.S. News & World Report in 2007.

All of the exam schools admit new students for grades 7 and 9, but in addition to this O'Bryant also admits some students for grade 10. In order to be admitted to one of the exam schools, a student is required to be a Boston resident. Students can apply to the exam schools from both inside and outside BPS. Each applicant submits a preference ordering of the exam schools to which they are applying. The admissions decisions are based on the applicants' Grade Point Average (GPA) in English and Math from the previous school year and the fall term of the ongoing school year as well as the applicants' scores on the Independent School Entrance Examination (ISEE) administered during the fall term of the ongoing school year. The ISEE is an entrance exam used by several selective schools in the United States. It consists of five sections: Reading Comprehension, Verbal Reasoning, Mathematics Achievement, Quantitative Reasoning, and a 30-minute essay. Exam school admissions only use the first four sections of the ISEE.

Each applicant receives an offer from at most one exam school, and waitlists are not used. The assignment of exam school offers is based on the student-proposing Deferred Acceptance (DA) algorithm by Gale and Shapley (1962). The algorithm takes as inputs each exam school's predetermined capacity, the applicants' preferences over the exam schools, and the exam schools' rankings of the applicants based on a weighted average of their standardized GPA and ISEE scores. These rankings differ slightly across the exam schools as for each school the standardization and ranking is done only within the pool of applicants to that school.

The DA algorithm produces exam school-specific admissions cutoffs that are given by the lowest rank among the applicants admitted to a given exam school. Since applicants receive an offer from at most one exam school, there is not a direct link between the exam school-specific running variables (an applicant's rank among applicants to a given exam school) and the exam school offers. However, as in Abdulkadiroglu, Angrist, and Pathak (2014), it is possible to construct a sharp sample for each exam school that consists of applicants who receive an offer if and only if their running variable is above the admissions cutoff for the exam school in question. Appendix B describes in detail the DA algorithm and the construction of the sharp samples.

1.4.2 Data

The main data for this paper comes from three sources provided by the BPS: (1) an exam school application file, (2) a BPS registration and demographic file, and (3) an MCAS file. These files can be merged together using a unique BPS student identification number. In addition, I use the students' home addresses to merge the BPS data with Census tract-level information from the American Community Survey (ACS) 5-year summary file for 2006-2011.[23]

The exam school application file consists of the records for all exam school applications in 1995-2009. It provides information on each applicant's application year and grade, application preferences, GPA in English and Math, ISEE scores, exam school-specific ranks, and the admissions decision. This allows me to reproduce the exam school-specific admissions cutoffs (the lowest rank among applicants admitted to a given

[23]See Abdulkadiroglu, Angrist, and Pathak (2014) for a more detailed description of the BPS data.

exam school). I transform the exam school-specific ranks into percentiles, ranging from 0 to 100, within application year and grade. I then center these running variables to be 0 at the admissions cutoff for the exam school in question. Thus, the running variables give an applicant's distance from the admissions cutoff in percentile units. Finally, I standardize the ISEE scores and GPA to have a mean of 0 and a standard deviation of 1 in the applicant population within each year and grade.[24]

The BPS registration and demographic file consists of the records for all BPS students in 1996-2012. It provides information on each student's home address, school, grade, gender, race, limited English proficiency (LEP) status, bilingual status, special education (SPED) status, and free or reduced price lunch (FRLP) status.

The MCAS file consists of the records for all MCAS tests taken by BPS students in 1997-2008. It provides information on 4th, 7th, and 10th grade MCAS scores in English, and 4th, 8th, and 10th grade MCAS scores in Math. In the case of retakes I only consider the first time a student took the test. I construct middle school and high school MCAS composites as the average MCAS scores in 7th grade English and 8th grade Math and 10th grade English and Math. I standardize the 4th grade MCAS scores in English and Math as well as the middle school and high school MCAS composite scores to have a mean 0 and a standard deviation of 1 in the BPS population within each year and grade.

Lastly, I use the ACS 5-year summary file for 2006-2011 to obtain information on the median family income, percent of households occupied by the owner, percent of families headed by a single parent, percent of households where a language other than English is spoken, the distribution of educational attainment, and the number of school-aged children in each Census tract in Boston. I use this information to divide the Census tracts into socioeconomic tiers as described in Section 1.6.1.

I restrict the sample to students who applied to the exam schools for 7th grade in 2000-2004. I focus on 7th grade applicants as most students enter the exam schools in 7th grade, and their exposure to the exam school treatment is longer. This is also the applicant group for which the covariate-based RD extrapolation approach by Angrist and Rokkanen (2013) fails. The restriction to application years 2000-2004 is done in order to have both 4th grade MCAS scores and middle/high school MCAS composite scores for the applicants. I exclude students who apply to the exam schools from outside BPS as these applicants are more likely to remain outside BPS and thus not have follow up information in the data. In addition, I exclude students with missing covariate or 4th grade MCAS score information.

Table 1.1 reports descriptive statistics for all BPS students as well as the exam school applicants in the estimation sample. Column (1) includes all BPS students enrolled in 6th grade in 2000-2004. Column (2) includes the subset of students who apply to the exam schools. Columns (3)-(6) include the subsets of applicants who receive no exam school offer or an offer from a given exam school. Exam school applicants are a highly selected group of students, with markedly higher 4th grade MCAS scores and lower shares of blacks and Hispanics, limited English proficiency, and special education than BPS students as a whole. Similarly,

[24]The exam school application file only contains a combined index of GPA in English and Math.

there is considerable selection even within exam school applicants according to their exam school assignment, with applicants admitted to a more selective exam school having higher 4th grade MCAS scores and lower shares of blacks and Hispanics, limited English proficiency, and students eligible for free or reduced price lunch.

1.4.3 Identification and Estimation

Throughout the rest of the paper I use $Z \in \{0, 1, 2, 3\}$ to denote the exam school assignment of an applicant where 0 stands for no offer, 1 for O'Bryant, 2 for Latin Academy and 3 for Latin School. Furthermore, I use $S \in \{0, 1, 2, 3\}$ to denote the enrollment decision of an applicant in the fall following exam school application where 0 stands for traditional Boston public school, 1 for O'Bryant, 2 for Latin Academy, and 3 for Latin School. Lastly, I use R_1, R_2, and R_3 to denote the running variables for O'Bryant, Latin Academy, and Latin School.

As discussed in Section 1.4.1, each applicant receives at most one exam school offer that is determined by the DA algorithm. The exam school assignment of an applicant is a deterministic function of her running variables and application preferences, denoted by P,

$$Z = g_Z(R_1, R_2, R_3, P).$$

The running variables are deterministic functions of the applicant's scores in the Reading Comprehension, Verbal Reasoning, Mathematics Achievement, and Quantitative Reasoning sections of the ISEE, denoted by M_2^E, M_3^E, M_2^M, and M_3^M, as well as her GPA in English and Math, denoted by G,

$$R_s = g_{R_s}\left(M_2^E, M_3^E, M_2^M, M_3^M, G\right), \; s = 1, 2, 3.$$

In addition, the data contains 4th grade MCAS scores in English and Math, denoted by M_1^E and M_1^M.

I treat the 4th grade MCAS score in English and the scores in the Reading Comprehension and Verbal Reasoning sections of the ISEE as noisy measures of an applicant's English ability, denoted by θ_E,

$$M_k^E = g_{M_k^E}\left(\theta_E, \nu_{M_k^E}\right), \; k = 1, 2, 3.$$

I treat the 4th grade MCAS score in Math and the scores in the Mathematics Achievement and Quantitative Reasoning sections of the ISEE instead as noisy measures of an applicant's Math ability, denoted by θ_M,

$$M_k^M = g_{M_k^M}\left(\theta_M, \nu_{M_k^M}\right), \; k = 1, 2, 3.$$

The test scores are strongly correlated with each other, but this correlation is far from perfect: some of the applicants scoring well in one test perform badly in another test. This can be seen from the scatterplots

and correlation in Figure 1-8 and Table 1.4. Consistent with the latent factor structure specified above, test scores measuring English ability are more highly correlated with each other than with test scores measuring Math ability, and vice versa. The only exception to this is 4th grade MCAS score in English that is most highly correlated with 4th grade MCAS score in Math. There is also a clear time-pattern among test scores measuring a given ability: the ISEE scores measuring the same ability are more highly correlated with each other than with the 4th grade MCAS score measuring the same ability.

Let $Y(s)$, $s = 0, 1, 2, 3$, denote potential outcomes under different enrollment decisions, and let $S(z)$, $z = 0, 1, 2, 3$, denote potential enrollment decisions under different exam school assignments. I assume that the potential outcomes and enrollment decisions are jointly independent of the test scores conditional on English and Math abilities and a set of covariates, denoted by X. Formally,

$$\left(\{Y(s)\}_{s=0}^{3}, \{S(z)\}_{z=0}^{3}\right) \perp\!\!\!\perp M \mid \theta, X,$$

where $M = \left(M_1^E, M_2^E, M_3^E, M_1^M, M_2^M, M_3^M\right)$. The covariates included in X are GPA, application preferences, application year, race, gender, SES tier as well as indicators for free or reduced price lunch, limited English proficiency, special education, and being bilingual. I also assume that there is sufficient noise in the ISEE scores so that conditional on the covariates it is possible to observe an applicant with a given level of English and Math ability under any exam school assignment. Formally, this common support assumption is given by

$$0 < P[Z = z \mid \theta, X] < 1, \; z = 0, 1, 2, 3.$$

Together these two assumptions can be used to identify causal effects of different exam school assignments on either enrollment or achievement, as discussed in Section 1.2. I make two additional assumptions that allow me to also identify causal effects of enrollment at a given exam school as opposed to a traditional Boston public school for the compliers who enroll at the exam school if they receive an offer and enroll at a traditional Boston public school if they receive no offer. First, I assume that receiving an offer from exam school s as opposed to no offer induces at least some applicants to enroll at exam school s instead of a traditional Boston public school. Second, I assume that this is the only way in which receiving an offer from exam school s as opposed to no offer can affect the enrollment decision of an applicant.[25] Formally, these first stage and monotonicity assumptions are given by

$$\begin{aligned} P[S(s) = s, S(0) = 0 \mid \theta, X] &> 0, \; s = 1, 2, 3 \\ P\left[S(s) = s', S(0) = s'' \mid \theta, X\right] &= 0, \; s' \neq s, s'' \neq 0. \end{aligned}$$

[25] I also assume in general that receiving an offer from exam school s can only induce an applicant to attend this school as opposed to another school. This rules out, for instance, the case in which an applicant is induced to enroll at Latin School as opposed to Latin Academy by receiving an offer from O'Bryant as opposed to no exam school offer.

In the estimation I approximate the conditional joint distribution of English and Math ability by a bivariate normal distribution,

$$\begin{bmatrix} \theta_E \\ \theta_M \end{bmatrix} \mid X \sim N\left(\begin{bmatrix} \mu_{\theta_E}' X \\ \mu_{\theta_M}' X \end{bmatrix}, \begin{bmatrix} \sigma^2_{\theta_E} & \sigma_{\theta_E\theta_M} \\ \sigma_{\theta_E\theta_M} & \sigma^2_{\theta_M} \end{bmatrix} \right).$$

To ensure a valid variance-covariance matrix I use the parametrization

$$\begin{bmatrix} \sigma^2_{\theta_E} & \sigma_{\theta_E\theta_M} \\ \sigma_{\theta_E\theta_M} & \sigma^2_{\theta_M} \end{bmatrix} = \begin{bmatrix} \omega_{11} & 0 \\ \omega_{21} & \omega_{22} \end{bmatrix} \begin{bmatrix} \omega_{11} & \omega_{21} \\ 0 & \omega_{22} \end{bmatrix}.$$

I also approximate the conditional distributions of the test scores using normal distributions given by

$$\begin{aligned} M_k^E \mid \theta, X &\sim N\left(\mu_{M_k^E}' X + \lambda_{M_k^E}\theta_E, \exp\left(\gamma_{M_k^E} + \delta_{M_k^E}\theta_E\right)^2 \right), \; k = 1,2,3 \\ M_k^M \mid \theta, X &\sim N\left(\mu_{M_k^M}' X + \lambda_{M_k^M}\theta_M, \exp\left(\gamma_{M_k^M} + \delta_{M_k^M}\theta_M\right)^2 \right), \; k = 1,2,3, \end{aligned}$$

where $\mu_{M_1^E} = \mu_{M_1^M} = 0$ and $\lambda_{M_1^E} = \lambda_{M_1^M} = 1$ to pin down the location and scale of the abilities as discussed in Section 1.2.3. Thus, I restrict the conditional expectations of the measurements to depend linearly on ability and allow for heteroskedasticity in the measurement error with respect to ability. Finally, I restrict the measurements to be jointly independent conditional on the abilities and covariates.

Let $D_s(z) = 1(S(z) = s)$, $s = 0,1,2,3$, denote indicators for potential enrollment decisions under different exam school assignments, and let $Y(S(z))$ denote potential outcomes under different exam school assignments. I approximate the conditional expectations of $D_s(z)$ and $Y(S(z))$ using the linear models

$$\begin{aligned} E[D_s(z) \mid \theta, X] &= \alpha_{D_s(z)}' X + \beta^E_{D_s(z)}\theta_E + \beta^M_{D_s(z)}\theta_M \\ E[Y(S(z)) \mid \theta, X] &= \alpha_{Y(S(z))}' X + \beta^E_{Y(S(z))}\theta_E + \beta^M_{Y(S(z))}\theta_M \end{aligned}$$

where $z = 0,1,2,3$.

The identification of the measurement and latent outcome models specified above follows directly from the nonparametric identification results presented in Sections 1.2 and 1.3. I illustrate this in more detail in Appendix C by providing moment equations that identify these particular parametric models.

I estimate the parameters of the measurement model using Maximum Simulated Likelihood (MSL). I use 500 random draws from the conditional joint distribution of θ given X, $f_{\theta|X}$, to evaluate the integral in the

conditional joint density of M given X, $f_{M|X}$. For a given observation $f_{M|X}$ is given by

$$\begin{aligned} & f_{M|X}\left(m \mid X;\mu,\lambda,\gamma,\delta,\omega\right) \\ &= \int \left[\prod_{k=1}^{3} f_{M_k^E|\theta,X}\left(m_k^E \mid \theta,X;\mu,\lambda,\gamma,\delta\right) f_{M_k^M|\theta,X}\left(m_k^M \mid \theta,X;\mu,\lambda,\gamma,\delta\right)\right] \\ & \quad \times f_{\theta|X}\left(\theta \mid X;\mu,\omega\right) d\theta \end{aligned}$$

where the conditional densities $f_{M_k^E|\theta,X}$, $f_{M_k^M|\theta,X}$, $k=1,2,3$, and $f_{\theta|X}$ are as specified above.

I estimate the parameters of the latent outcome models using the Method of Simulated Moments (MSM) based on the moment equations

$$\begin{aligned} E\left[D_s \mid M,X,Z\right] &= \alpha_{D_s(Z)}' X + \beta_{D_s(Z)}^E E\left[\theta_E \mid M,X,Z\right] + \beta_{D_s(Z)}^M E\left[\theta_M \mid M,X,Z\right] \\ E\left[Y \mid M,X,Z\right] &= \alpha_{Y(S(z))}' X + \beta_{Y(S(z))}^E E\left[\theta_E \mid M,X,Z\right] + \beta_{Y(S(z))}^M E\left[\theta_M \mid M,X,Z\right] \end{aligned}$$

for $Z=0,1,2,3$. The conditional expectations $E\left[\theta_E \mid M,X,Z\right]$ and $E\left[\theta_M \mid M,X,Z\right]$ are computed using the MSL estimates of the measurement model and 500 random draws from $f_{\theta|X}$. The weighting matrix in the MSM procedure is based on the number of observations in the (M,X,Z) cells. This implies that the parameters of the latent outcome models can be estimated in practice by running a regression of D_S or Y on X, $E\left[\theta_E \mid M,X,Z\right]$, and $E\left[\theta_M \mid M,X,Z\right]$ using observations with $Z=0,1,2,3$.

The standard errors presented below are based on nonparametric 5-step bootstrap using 500 replications (Davidson and MacKinnon, 1999; Andrews, 2002). For each bootstrap sample I re-estimate the measurement model using the original estimates as initial values and stop the MSL procedure after five iterations. I then re-estimate the latent outcome models using these MSL estimates. This provides a computationally attactive approach for taking into account the uncertainty related to both step of the estimation procedure due to the slow speed of convergence of the MSL estimation.

1.5 Extrapolation Results

1.5.1 Effects at the Admissions Cutoffs

To benchmark the latent factor model-based estimates, I begin with RD estimates of causal effects of exam school attendance for marginal applicants at the admissions cutoffs in the sharp samples. Figures 1-6a and 1-6b plot the relationship between the running variables and the probabilities of receiving an offer from and enrolling at a given exam school in windows of ± 20 around the admissions cutoffs in the sharp samples. The blue dots show bin averages in windows of width 1. The black solid lines show fits from local linear regressions estimated separately to the left and right of the cutoffs using the edge kernel and a bandwidth computed separately for each exam school using the algorithm by Imbens and Kalyanaraman (2012). Figures

1-7a and 1-7b show the same plots for average middle school and high school MCAS composite scores.

Table 1.2 reports the first stage, reduced form, and Local Average Treatment Effect estimates corresponding to Figures 1-6 and 1-7. The estimates are based on local linear regressions using the edge kernel and a bandwidth that is computed separately for each exam school and MCAS outcome using the algorithm by Imbens and Kalyanaraman (2012).[26] The first stage and reduced form models are given by

$$\begin{aligned} D_s &= \alpha_{FS} + \beta_{FS} Z_s + \gamma_{FS} R_s + \delta_{FS} Z_s \times R_s + X^{'} \pi_{FS} + \eta \\ Y &= \alpha_{RF} + \beta_{RF} Z_s + \gamma_{RF} R_s + \delta_{RF} Z_s \times R_s + X^{'} \pi_{RF} + \epsilon \end{aligned}$$

where D_s is an indicator for enrollment at exam school s in the following school year, Y is the outcome of interest, Z_s is an indicator for being at or above the admissions cutoff for exam school s, R_s is the distance from admissions cutoff for exam school s, and X is a vector containing indicators for application years and application preferences. The first stage and reduced form estimates are given by β_{FS} and β_{RF}, and the Local Average Treatment Effect estimate is given by the ratio $\frac{\beta_{RF}}{\beta_{FS}}$. In practice this ratio can be estimated using weighted 2-Stage Least Squares (2SLS).

Figure 1-6a confirms the sharpness of exam school offers as functions of the running variables in the sharp samples discussed in Section 1.4.1: the probability of receiving an offer from a given exam school jump from 0 to 1 at the admissions cutoff. However, as can be seen from Figure 1-6b and the first stage estimates in Table 1.2, not all applicants receiving an offer from a given exam school choose to enroll there. The enrollment first stages are nevertheless large. An offer from O'Bryant raises the probability of enrollment at O'Bryant from 0 to .78 at the admissions cutoff whereas offers from Latin Academy and Latin School raise the probability of enrollment at these schools from 0 to .95 and .96.

Exam school offers have little effect on the average middle school and high school MCAS composite scores of the marginal applicants, as can be seen from Figures 1-7a and 1-7b and the reduced form estimates in Table 1.2. The only statistically significant effect is found for middle school MCAS composite score at the Latin Academy admissions cutoff: an offer from Latin Academy is estimated to reduce the average score by $.181\sigma$. According to the corresponding Local Average Treatment Effect estimate in Table 1.2, enrolling at Latin Academy leads to a $.191\sigma$ reduction in the average score among compliers at the admissions cutoff.

Table 1.3 repeats the estimations separately for applicants whose average 4th grade MCAS scores fall below and above the within-year median. The first stage estimates show large enrollment effects at the admissions cutoffs for both lower-achieving and higher-achieving applicants. These effects are similar in magnitude to the effects estimated for the full sample. The reduced form and Local Average Treatment Effect estimates for applicants with low average 4th grade MCAS scores are relatively noisy due to small sample size, but there is some evidence of treatment effect heterogeneity by prior achievement, a point I return to in Section 1.5.2. The reduced form estimate suggests that an offer from O'Bryant increases average

[26] As noted by Calonico, Cattaneo, and Titiunik (2014), the algorithm by Imbens and Kalyanaraman (2012) may generate too large bandwidths. However, my findings are not sensitive to alternative ways of choosing the bandwidths.

high school MCAS composite score by .204σ at the admissions cutoff among applicants with low average 4th grade MCAS scores. The corresponding Local Average Treatment Effect estimate suggests that enrolling at O'Bryant increases the average score among the compliers at the admissions cutoff by .275σ. The reduced form and Local Average Treatment Effect estimates for applicants with high average 4th grade MCAS scores are similar to the estimates for the full sample.

It is important to note the incremental nature of the RD estimates reported above. Applicants just below the O'Bryant admissions cutoff do not receive an offer from any exam school, meaning that the counterfactual for these applicants is a traditional Boston public school. On the other hand, the vast majority of applicants just below the Latin Academy admissions cutoff receive an offer from O'Bryant, and the vast majority of applicants just below the Latin School admissions cutoff receive an offer from Latin Academy. Thus, the reduced form and Local Average Treatment Effect estimates for Latin Academy and Latin School should be interpreted as the effect of receiving an offer from and enrolling at a more selective exam school. I return to this point in Section 1.5.3.

1.5.2 Estimates of the Latent Factor Model

Before investigating effects away from the admissions cutoffs I briefly discuss the main estimates of the latent factor model. Figure 1-9 shows the underlying marginal distributions of English and Math ability in the population of exam school applicants. I have constructed these distributions using kernel density estimates based on simulations from the estimated measurement model. The marginal distributions look relatively normal which is expected given the joint normality assumption on the conditional distribution of the abilities given covariates. The mean and standard deviation of English ability are 1.165 and .687. The mean and standard deviation of Math ability are 1.121 and .831. Figure 1-10 shows a scatterplot of English and Math abilities. The relationship between the abilities is relatively linear which is expected given the joint normality assumption on the conditional distribution. The correlation between English and Math ability is .817.

Table 1.5 reports estimates of the factor loadings on the means and (log) standard deviations of the measures. As discussed in Section 1.4, I pin down the scales of the abilities by normalizing the factor loadings on the means of 4th grade MCAS scores to 1. The estimated factor loadings on the means of ISEE scores are instead slightly above 1. The estimated factor loadings on the (log) standard deviations of ISEE scores in Reading Comprehension and Verbal Reasoning and 4th grade MCAS score in Math suggest that the variances of these measures are increasing in ability. The estimated factor loadings on the (log) standard deviations of ISEE scores in Mathematical Achievement and Quantitative Reasoning and 4th grade MCAS score in English are small and statistically insignificant.

Table 1.6 reports estimates of the factor loadings on enrollment (First Stage) and middle school and high school MCAS composite scores (Reduced Form) under a given exam school assignment. For no exam school offer in Column (1) the enrollment outcome is enrollment at a traditional Boston public school. For an offer

from a given exam school in Columns (2), (3), and (4) the outcome is enrollment at the exam school in question. The estimated factor loadings on enrollment are largely negative, suggesting that applicants with higher ability are less likely to attend a given exam school if they receive an offer (or a traditional Boston public school if they receive no offer). However, the opposite is true for Latin School.

Unlike for enrollment, the estimated factor loadings on middle school and high school MCAS composite scores are positive, large in magnitude, and higly statistically significant. This is not surprising: applicants with higher English and Math abilities perform, on average, better in middle school and high school MCAS exams in English and Math irrespective of their exam school assignment. A more interesting finding arising from these estimates is that the factor loadings tend to be larger in magnitude under no exam school offer than under an offer from any given exam school. This is especially true for high school MCAS composite scores. This suggest that applicants with lower English and Math abilities benefit more from access to exam schools. I return to this point below.

1.5.3 Effects Away from the Admissions Cutoffs

The estimates of the latent factor model can be used to construct empirical counterparts of Figures 1-1 and 1-5 that illustrate the extrapolation problem in sharp RD and the latent factor-based approach to solving this problem. Figure 1-11 plots the latent factor model-based fits and extrapolations of the potential outcomes in the RD experiments for the sharp samples over the full supports of the running variables. The blue dots show bin averages in windows of width 1. The black solid lines show the latent factor model-based fits, and the dashed red lines show the latent factor model-based extrapolations. I smooth the fits and extrapolations with a local linear regression using the edge kernel and a rule of thumb bandwidth (Fan and Gijbels, 1996). The fits and extrapolations are defined as

$$E\left[Y\left(S\left(z^{act}\right)\right) \mid R_s = r\right],\ s = 1, 2, 3$$
$$E\left[Y\left(S\left(z^{cf}\right)\right) \mid R_s = r\right],\ s = 1, 2, 3$$

where the counterfactual assignment z^{cf} is an offer from exam school s for applicants below the admissions cutoff and an offer from the next most selective exam school for the applicants above the admissions cutoff (no exam school offer in the case of O'Bryant).

Figure 1-11a plots the fits and extrapolations for middle school MCAS composite scores. For O'Bryant the fits and extrapolations lie on top of each other, suggesting that receiving an offer from O'Bryant has no effect on the average score of either marginal applicants at the admissions cutoff or inframarginal applicants away from the admissions cutoff. Similarly, the extrapolations for Latin Academy reveal that the negative effect of receiving an offer from Latin Academy found for marginal applicants at the admissions cutoff in Section 1.5.1 holds also for inframarginal applicants away from the admissions cutoff. For Latin School the picture arising is markedly different. The extrapolations suggest that receiving an offer from Latin School has no

effect on the average score of inframarginal applicants above the admissions cutoff. Inframarginal applicants below the admissions cutoff would instead experience, on average, achievement gains from receiving an offer from Latin School.

Figure 1-11b plots the fits and extrapolations for high school MCAS composite scores. For all three exam schools the extrapolations suggest little effect from receiving an offer for inframarginal applicants above the admissions cutoffs, with the exception of applicants far above the O'Bryant admissions cutoff for which the effect is negative. For inframarginal applicants below the admissions cutoffs the picture arising is instead markedly different. For all three exam schools the extrapolations suggest a positive effect from receiving an offer for applicants failing to gain access to the exam school in question.

Table 1.7 reports estimates of the extrapolated first stage and reduced form effects of an offer from a given exam school on enrollment and middle school and high school MCAS composite scores. In addition, the table reports estimates of the extrapolated Local Average Treatment Effects of enrolling at a given exam school on middle school and high school MCAS composite scores for the compliers. The estimates are for the full population of applicants in the sharp samples. The estimates for middle school MCAS composite scores show no effect of an offer from or enrolling at O'Bryant. Offer from Latin Academy and Latin school are instead estimated to reduce the average score by $.229\sigma$ and $.214\sigma$. The corresponding Local Average Treatment Effect estimates are 0.236σ for Latin Academy and $.0226\sigma$ for Latin School. The estimates for high school MCAS composite scores show large positive effects for all three exam schools. The reduced form and Local Average Treatment Effect estimates are $.252\sigma$ and $.293\sigma$ for O'Bryant, $.279\sigma$ and $.290\sigma$ for Latin Academy, and $.199\sigma$ and $.209\sigma$ for Latin School.

Table 1.8 reports the same estimates separately for applicants below and above the admissions cutoffs. The first stage estimates reveal that the effects of an offer from a given exam school on enrollment at this school are larger among inframarginal applicants below the admissions cutoffs than among inframarginal applicants above the admissions cutoffs. The reduced form and Local Average Treatment Effect estimates for middle school MCAS composite scores show similar negative effects of an offer from and enrolling at Latin Academy as in Table 1.7 both below and above the admissions cutoff. The negative Latin School effect reported above is instead entirely driven by inframarginal applicants below the admissions cutoff. The reduced form and Local Average Treatment Effect estimates for high school MCAS composite scores confirm the above findings of large positive effects on the average score of inframarginal applicants below the admissions cutoffs. There is instead little evidence of effects for inframarginal applicants above the admissions cutoffs.

Similar to the RD estimates at the admissions cutoffs discussed in Section 1.5.1, the above estimates should be interpreted as incremental effects of receiving an offer form or enrolling at a more selective exam school. Thus, these estimates leave unanswered the question of how receiving an offer from or enrolling at a given exam school versus a traditional Boston public school affects achievement. Table 1.9 addresses this question by reporting the estimates of the extrapolated first stage and reduced form effects of receiving

an offer from a given exam school versus no offer from any exam school on enrollment at the exam school in question on middle school and high school MCAS composite scores. In addition, the table reports the extrapolated Local Average Treatment Effect estimates of the effect of enrolling at a given exam school versus a traditional Boston public school on middle school and high school MCAS composite scores for the compliers in the full population of applicants.

According to the estimates for middle school MCAS composite scores, offers from Latin Academy and Latin School versus no exam school offer reduce the average score by $.275\sigma$ and $.319\sigma$. The corresponding Local Average Treatment Effects of enrolling at a given exam school versus a traditional Boston public school are $-.288\sigma$ for Latin Academy, and $-.330\sigma$ for Latin School. There is instead no evidence of effects for O'Bryant. The estimates for high school MCAS composite scores are small in magnitude and statistically insignificant.

Table 1.10 reports the same estimates separately for applicant who receive no exam school offer and for applicants who receive an exam school offer. The estimates for middle school MCAS composite scores show similar negative effects of an offer from and enrolling at Latin Academy and Latin School as in Table 1.9 among both applicant groups. The estimates for high school MCAS scores reveal instead substantial heterogeneity in the treatment effects. The estimates suggest that receiving an offer from a given exam school versus no offer from any exam school has large positive effects among lower-achieving applicants failing to gain access to the exam schools. The reduced from effects are $.334\sigma$ for O'Bryant, $.429\sigma$ for Latin Academy, and $.428\sigma$ for Latin School. The corresponding Local Average Treatment Effects of enrolling at a given exam school versus a traditional Boston public school are $.376\sigma$ for O'Bryant, $.435\sigma$ for Latin Academy, and $.448\sigma$ for Latin School. The estimates for higher-achieving applicants gaining access to the exam schools are instead negative and large in magnitude. The reduced from effects are $-.268\sigma$ for O'Bryant, $-.300\sigma$ for Latin Academy, and $-.348\sigma$ for Latin School. The corresponding Local Average Treatment Effects are $-.353\sigma$ for O'Bryant, $-.325\sigma$ for Latin Academy, and $-.353\sigma$ for Latin School.

1.5.4 Placebo Experiments

A natural concern regarding the results in the previous section is that they are just an artifact of extrapolations away from the admissions cutoffs. To address this concern, I study the performance of the model using a set of placebo experiments. I start by dividing the applicants receiving a given exam school assignment $z = 0, 1, 2, 3$ in half based on the within-year median of the running variable distribution for this population.[27] I re-estimate the latent outcome models to the left and right of the placebo cutoffs and use the resulting estimates to extrapolate away from these cutoffs. All of the applicants both to the left and to the right of the cutoffs in these placebo RD experiments receive the same exam school assignment. Thus, the extrapolations should show no effects if the identifying assumptions are valid and the empirical specifications provide reasonable approximations of the underlying data generating process.

[27]For applicant receiving no offer I use the average of their exam school-specific running variables.

Figure 1-12 plots the latent factor model-based fits and extrapolations in the placebo RD experiments.[28] Figure 1-12a plots the estimates for middle school MCAS composite scores, and Figure 1-12b plots the estimates for high school MCAS composite scores. The blue dots show bin averages in windows of width 1. The black solid lines show the latent factor model-based fits to the data. The dashed red lines show the latent factor model-based extrapolations. I smooth the fits and extrapolations with a local linear regression using the edge kernel and a rule of thumb bandwidth (Fan and Gijbels, 1996). For both outcomes and for each exam school assignment the fits and extrapolations lie on top of each other, thus providing evidence supporting the identifying assumptions and empirical specifications. The only notable exceptions to this can be seen for high school MCAS composite scores far below the placebo cutoff for applicants receiving no offer from any exam school and far above the placebo cutoff for applicants receiving an offer from O'Bryant.

Table 1.11 reports estimates of the placebo reduced form effects on middle school and high school MCAS composite scores. The estimates are shown for all applicants as well as separately for applicants below and above the placebo cutoffs. The estimated effects are small in magnitude and statistically insignificant, thus providing further support for the validity of the results presented in Section 1.5.3.

1.6 Counterfactual Simulations

1.6.1 Description of the Admissions Reforms

Estimates of treatment effects away from the exam school admissions cutoffs are useful for predicting effects of reforms that change the exam school assignments of inframarginal applicants. A highly contentious example of this is the use of affirmative action in exam school admissions. I use the estimates of the latent factor model to predict how two particular affirmative action reforms would affect the achievement of exam school applicants.

The first reform reintroduces in the admissions process minority preferences that were in place in the Boston exam school admissions in 1975-1998. In this counterfactual admissions process 65% of the exam school seats are assigned purely based on achievement. The remaining 35% of the exam school seats are reserved for black and Hispanic applicants and assigned based on achievement. The assignment of seats within each group is based on the DA algorithm discussed in Section 1.4.1.

The second reform introduces in the admissions process socioeconomic preferences that have been in place in the Chicago exam school admissions since 2010. In this counterfactual admissions process 30% of the exam school seats are assigned purely based on achievement. The remaining 70% of the exam school seats are divided equally across four sosiocconomic tiers and assigned within them based on achievement. The assignment of the seats within each group is again based on the DA algorithm.

I generate the socioeconomic tiers by computing for each Census tract in Boston a socioeconomic index

[28] I transform the running variables into percentile ranks within each year in the placebo RD experiments and re-centered them to be 0 at the placebo cutoff for expositional purposes.

that takes into account the following five characteristics: (1) median family income, (2) percent of households occcupied by the owner, (3) percent of families headed by a single parent, (4) percent of households where a language other than English is spoken, and (5) an educational attainment score.[29] The socioeconomic index for a given Census tract is given by the sum of its percentile ranks in each five characteristics among the Census tracts in Boston (for single-parent and non-English speaking households 1 minus the percentile rank is used). I assign each BPS student a socioeconomic index based on the Census tract they live in and divide the students into socioecomic tiers based on the quartiles of the socioeconomic index distribution in the BPS population within each year.

To study the effects of the two reforms I reassign the exam school offers based on the counterfactual admissions processes, considering only applicants in the estimation sample described in Section 1.4.2. I use as the capacity of a given exam school in a given year the number of offers it made to the applicants in the estimation sample in that year. The latent factor model then allows me to predict average middle school and high school MCAS composite scores based on the reassigned exam school offers.[30]

An important feature of both of the reforms is that they cause substantial changes to the admissions cutoffs faced by the exam school applicants. This means that if there is considerable treatment effect heterogeneity in terms of the running variables, predictions of the effects of the reforms based on treatment effects at admissions cutoffs are likely to be misleading. Based on the results in Section 1.5, this is the case for Boston exam schools. Thus, it is a first-order issue to take this heterogeneity into account when predicting the effects of the reforms.

As with all counterfactuals, there are other dimensions that may chance as a result of the reforms. First, the reforms potentially affect the composition of the pool of exam school applicants as some student face a decrease and some students an increase in their ex ante expected probability of being admitted to a given exam school.[31] Second, the reforms will lead to changes in the composition of applicants who are admitted and consequently enroll at the exam schools. These changes may affect the sorting of teachers across schools (Jackson, 2009) and the way teaching is targeted (Duflo, Dupas, and Kremer, 2011) as well as affect achievement directly through peer effects (Epple and Romano, 2011; Sacerdote, 2011).

However, the above discussion should not be seen as a concern regarding the latent factor-based extrapolation approach per se. It is possible to address the above caveats by building a richer model that incorporates these channels into the latent factor framework. For instance, one can build a model of the exam school application behavior of BPS students along the lines of the work by Walters (2013) on Boston charter schools and incorporate this into the latent factor framework. I leave this and other potential extensions for future

[29] The educational attainment score is calculated based on the educational attainment distribution among individuals over the age of 25: $educational\ attainment\ score = 0.2 \times (\%\ less\ than\ high\ school\ diploma) + 0.4 \times (\%\ high\ school\ diploma) + 0.6 \times (\%\ some\ college) + 0.8 \times (\%\ bachelors\ degree) + 1.0 \times (\%\ advanced\ degree)$.

[30] This exercise is closely related to the literatute on evaluating the effects of reallocations on the distribution of outcomes. See, for instance, Graham (2011), Graham, Imbens, and Ridder (2010), and Graham, Imbens, and Ridder (2013).

[31] For instance, Long (2004) and Andrews, Ranchhod, and Sathy (2010) find the college application behavior of high school students in California and Texas to be resposive to affirmative action and other targeted recruiting programs. However, the evidence on this is somewhat mixed (Card and Krueger, 2005; Antonovics and Backes, 2013).

research as they are outside the scope of this paper.

1.6.2 Simulation Results

The introduction of either minority or socioeconomic preferences substantially affects exam school assigments: $27 - 35\%$ of applicants are affected by the reforms. This can be seen from Table 1.12, which reports the actual and counterfactual exam school assignments under the two reforms. This is also evident in Table 1.13 that reports the admissions cutoffs faced by different applicant groups under the counterfactual admissions process. The counterfactual admissions cutoffs are expressed as distances from the actual admissions cutoffs. Minority applicants would face substantially lower admissions cutoffs under minority preferences than under the current admissions process whereas the opposite is true for non-minority applicants. Similarly, applicants from lower socioeconomic tiers would face lower admissions cutoffs under socioeconomic preferences than under the current admissions process.

Table 1.14 reports descriptive statistics for the exam school applicants based on their counterfactual assignments under minority and socioeconomic preferences. The most notable compositional changes caused by the two reforms can be seen among applicants receiving an offer from Latin School. Under both counterfactual admissions processes, Latin School admits students with considerably lower average 4th grade MCAS scores in English and Math. Similarly, the share of blacks and Hispanics among the admitted students to Latin School would more than double under minority preferences and close to double under socioeconomic preferences. Furthermore, the average 4th grade MCAS scores in Math and English are higher and the shares of blacks and Hispanics lower among the applicants receiving no offer from any exam school under both counterfactual reforms. Changes in the composition of applicants receiving offers from O'Bryant and Latin Academy are instead less marked.

I use the estimated latent factor model to predict potential outcomes for the exam school applicants under the counterfactual admissions processes. These predictions can be used to evaluate whether the changes in exam school assignments caused by the two reforms translate into effects on achievement. To answer this question, Table 1.15 reports Average Reassignment Effects (ARE) of the reforms on middle school and high school MCAS composite scores. The Average Reassignment Effect is given by the difference in average potential outcomes among the exam school applicants under the counterfactual and actual admissions processes:

$$\begin{aligned} E\left[Y^{cf} - Y^{act}\right] &= \sum_{z=0}^{3} P\left[Z^{cf} = z\right] E\left[Y\left(S\left(z\right)\right) \mid Z^{cf} = z\right] \\ &\quad - \sum_{s=0}^{3} P\left[Z^{act} = z\right] E\left[Y\left(S\left(z\right)\right) \mid Z^{act} = z\right] \end{aligned}$$

where Z^{act} and Z^{cf} are an applicant's actual and counterfactual exam school assignments. The table reports estimates both for the full population of applicants and for applicants whose exam school assignment

is affected by the reforms.

The introduction of minority preferences would have no effect on the average middle school MCAS composite score among exam school applicants. However, this masks substantial heterogeneity in the effects across minority and non-minority applicants. The estimates suggest that the reform would reduce the average score among minority applicants by $.028\sigma$ and increase it among non-minority applicants by $.043\sigma$. The estimated effects are larger among the affected applicants: $-.084\sigma$ for minority applicants and $.113\sigma$ for non-minority applicants. The estimates for high school MCAS composite scores suggest that the introduction of minority preferences would increase the average score by $.023\sigma$ among all applicants and by $.062\sigma$ among affected applicants. There is less marked heterogeneity in these effects across minority and non-minority applicants, but the effects are somewhat larger for minority applicants.

The introduction of socioeconomic preferences would have no effect on the average middle school MCAS composite score among exam school applicants. However, there is considerable heterogeneity in the effects across applicants from different socioeconomic tiers. The estimates suggest that the reform would reduce the average score among applicants from the lowest socioeconomic tier by $.032\sigma$ and increase the average score among applicants from the highest socioeconomic tier by $.041\sigma$. The estimated effects are larger among the affected applicants: $-.092\sigma$ for the lowest socioeconomic tier and $.133\sigma$ for the highest socioeconomic tier. The estimates for high school MCAS composite scores suggest that the introduction of socioeconomic preferences would increase the average score by $.015\sigma$ among all applicants and by $.050\sigma$ among affected applicants. There is again considerable heterogeneity in the effects across applicants from different socioeconomic tiers. The reform would increase the average score by $.068\sigma$ among applicants from the lowest socioeconomic tier and by $.054\sigma$ among applicants from the highest socioeconomic tier.

There are two mechanisms at work behind these estimates. First, the reforms lower the admissions cutoffs faced by minority applicants and applicants from lower socioeconomic tiers. This leads to more lower-achieving applicants, who experience achievement gains from exam school attendence, to gain access to the exam schools. Second, the reforms increase the admissions cutoff faced by non-minority applicants and applicants from higher socioeconomic tiers. This leads to some of the higher-achieving applicants, who experience achievement losses from exam school attendance, to lose their exam school seats.

1.7 Conclusions

RD design allows for nonparametric identification and estimation of treatment effects for individuals at the cutoff value determining treatment assignment. However, many policies of interest change treatment assignment of individuals away from the cutoff, making knowledge of treatment effects for these individuals of substantial interest. A highly contentious example of this is affirmative action in selective schools that affects admissions cutoffs faced by different applicant groups.

The contributions of this paper are two-fold. First, I develop a new latent factor-based approach to the

identification and estimation of treatment effects away from the cutoff in RD. The approach relies on the assumption that sources of omitted variables bias in an RD design can be modeled using unobserved latent factors. My main result is nonparametric identification of treatment effects for all values of the running variable based on the availability of multiple noisy measures of the latent factors. Second, I use the latent factor framework to estimate causal effects of Boston exam school attendance for the full population of applicants and to simulate effects of introducing either minority or socioeconomic preferences in exam school admissions.

My findings highlight the local nature of RD estimates that show little evidence of causal effects for marginal applicants at admissions cutoffs (Abdulkadiroglu, Angrist, and Pathak, 2014). The estimates of the latent factor model suggest that achievement gains from exam school attendance are larger among applicants with lower baseline measures of ability. As a result, lower-achieving applicants who currently fail to gain admission to Boston exam schools would experience substantial achievement gains from attending these schools. The simulations predict that the introduction of either minority or socioeconomic preferences in exam school admissions boosts average achievement among applicants. This is largely driven by achievement gains experienced by lower-achieving applicants who gain access to exam schools as a result of the policy change. These findings are of significant policy-relevance given ongoing discussion about the use of affirmative action in exam school admissions.

I focus in this paper on the heterogeneity in causal effects of exam school attendance based on the running variables used in the admissions process. This is a first-order concern when predicting effects of admissions reforms that widely change the exam school assignments of inframarginal applicants. However, as with all counterfactuals, there are other dimension that may change as a result of these reforms. First, affirmative action might lead to changes in the application behavior of students (Long, 2004; Andrews, Ranchhod, and Sathy, 2010). Second, affirmative action causes changes in student composition that may affect the sorting of teachers across schools (Jackson, 2009) as well as the way teaching is targeted (Duflo, Dupas, and Kremer, 2011). Finally, the changes in student composition may affect achievement directly through peer effects (Epple and Romano, 2011; Sacerdote, 2011). It is possible to model these channels in the latent factor framework, but this is left for future research.

Boston exam schools, as well as other selective schools, are a natural application for latent factor-based RD extrapolation as admissions are based on noisy measures of applicants' latent abilities. However, the approach is likely to prove useful also in other educational settings, such as gifted and talented programs (Bui, Craig, and Imberman, forthcoming) and remedial education (Jacob and Lefgren, 2004; Matsudaira, 2008). Moreover, the approach is likely to prove useful in health settings where treatment assignment is based on noisy measures of individuals' latent health conditions. Such settings include, for instance, the use of birth weight to assign additional medical care for newborns (Almond, Doyle, Kowalski, and Williams, 2010; Bharadwaj, Loken, and Neilson, 2013). As illustrated by the findings for Boston exam schools, local effects identified by RD do not necessarily represent the effects of policy interest. Latent factor-based RD

extrapolation provides a framework for investigating external validity in these and other RD designs.

1.8 Figures and Tables

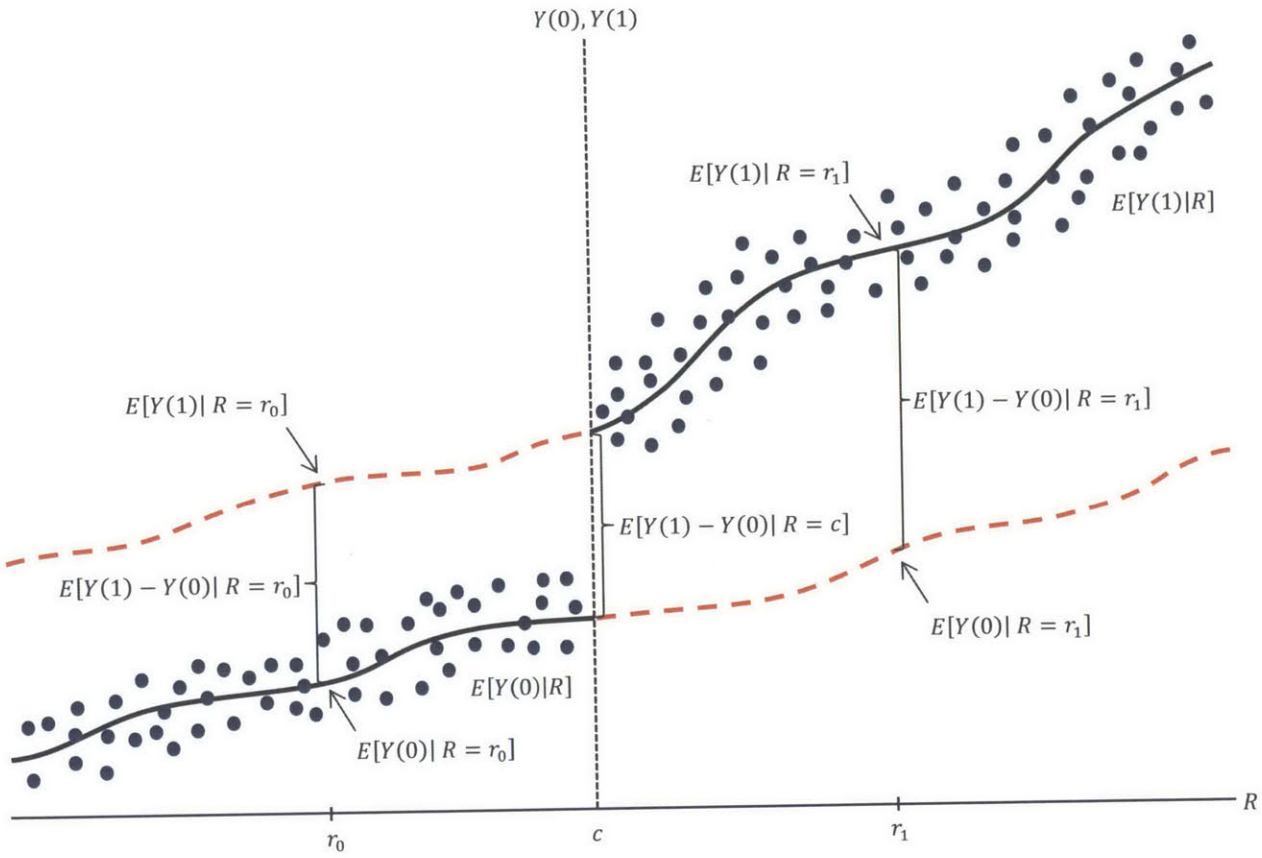


Figure 1-1: Extrapolation Problem in a Sharp Regression Discontinuity Design

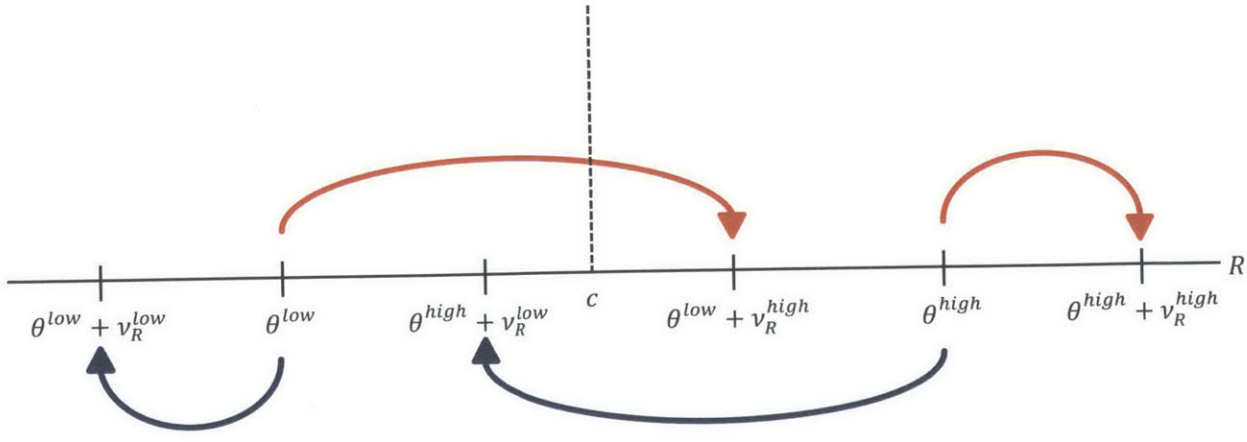


Figure 1-2: Treatment Assignment in a Latent Factor Framework

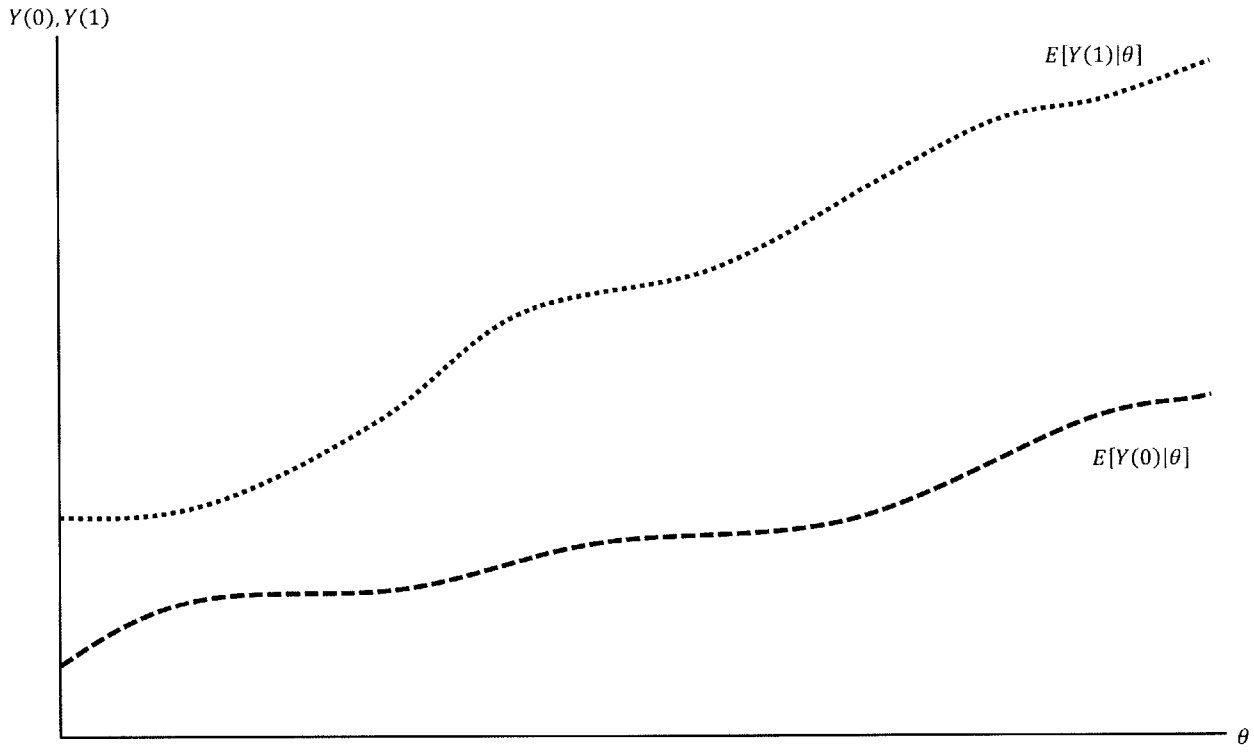


Figure 1-3: Latent Conditional Expectation Functions

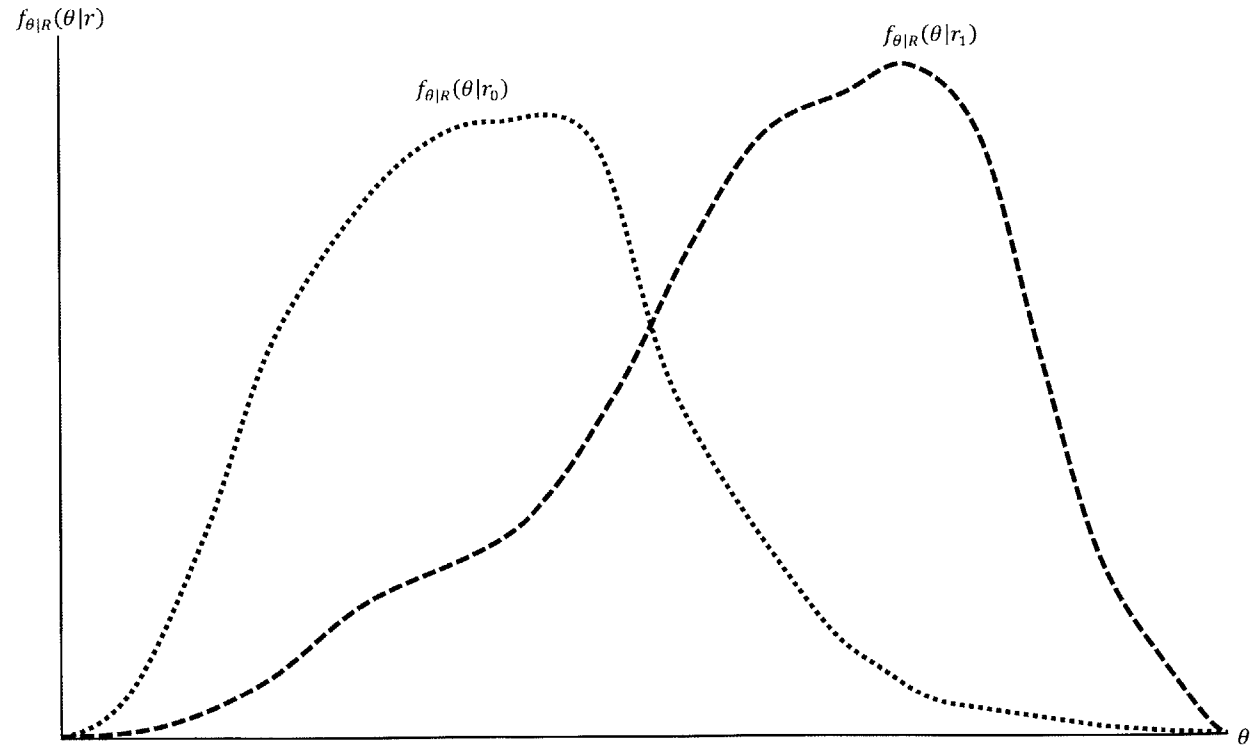


Figure 1-4: Conditional Latent Factor Distributions Given the Running Variable

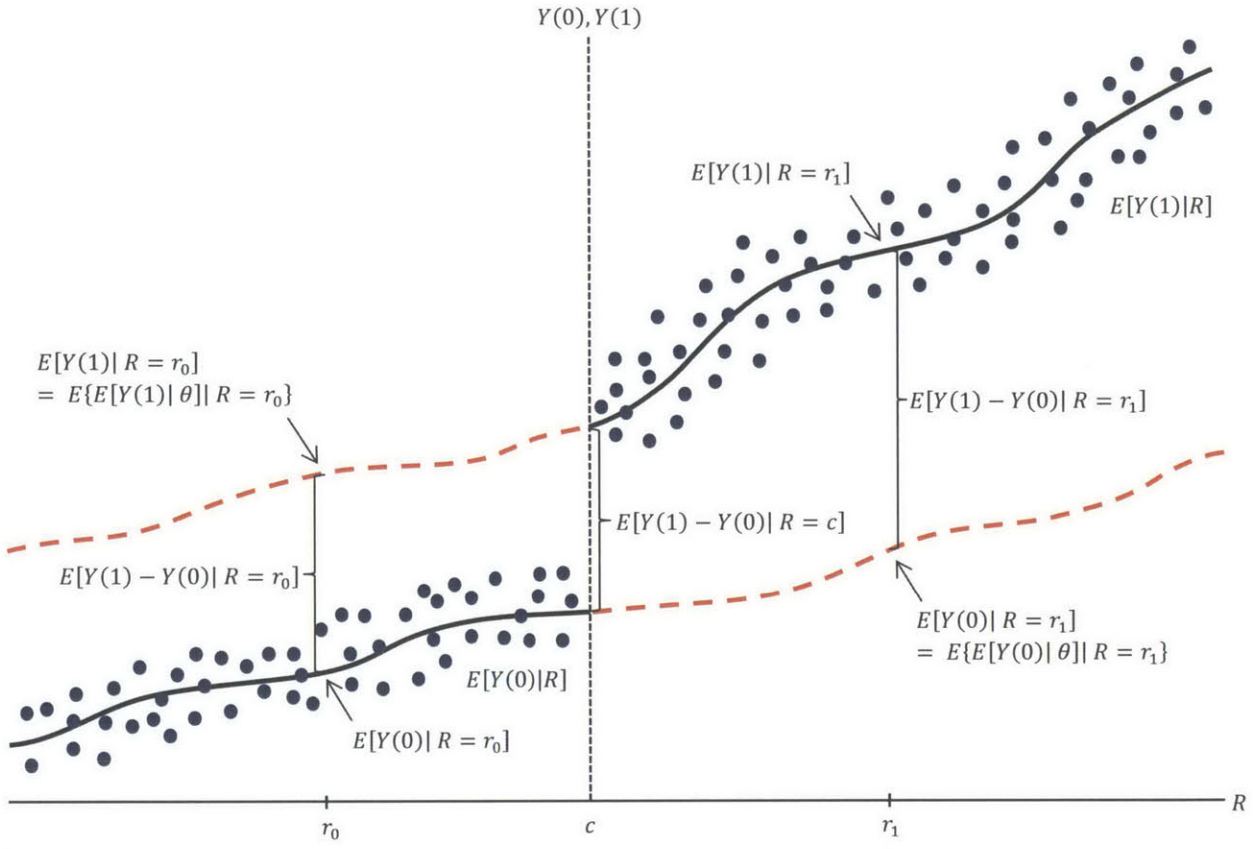
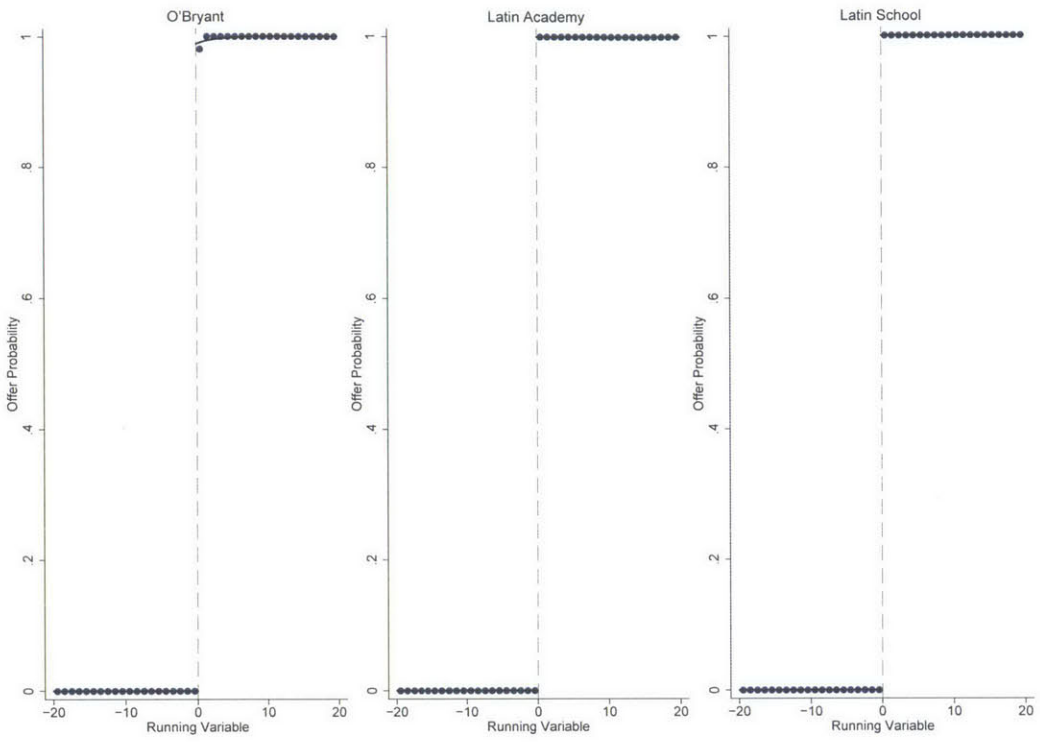
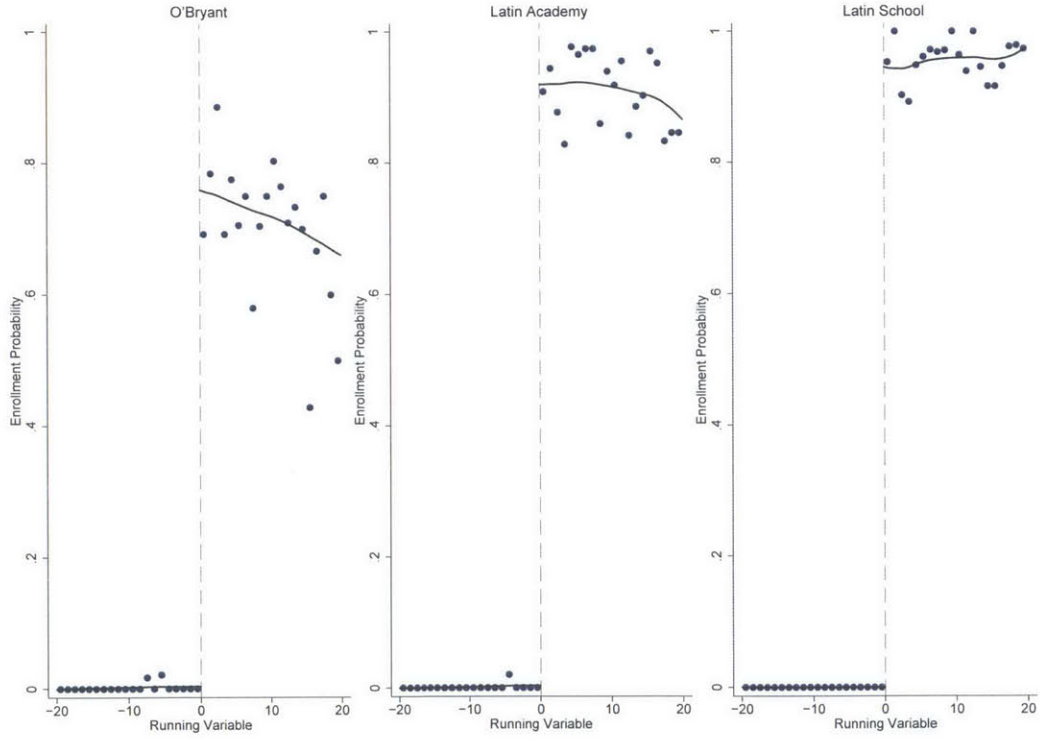


Figure 1-5: Latent Factor-Based Extrapolation in a Sharp Regression Discontinuity Design

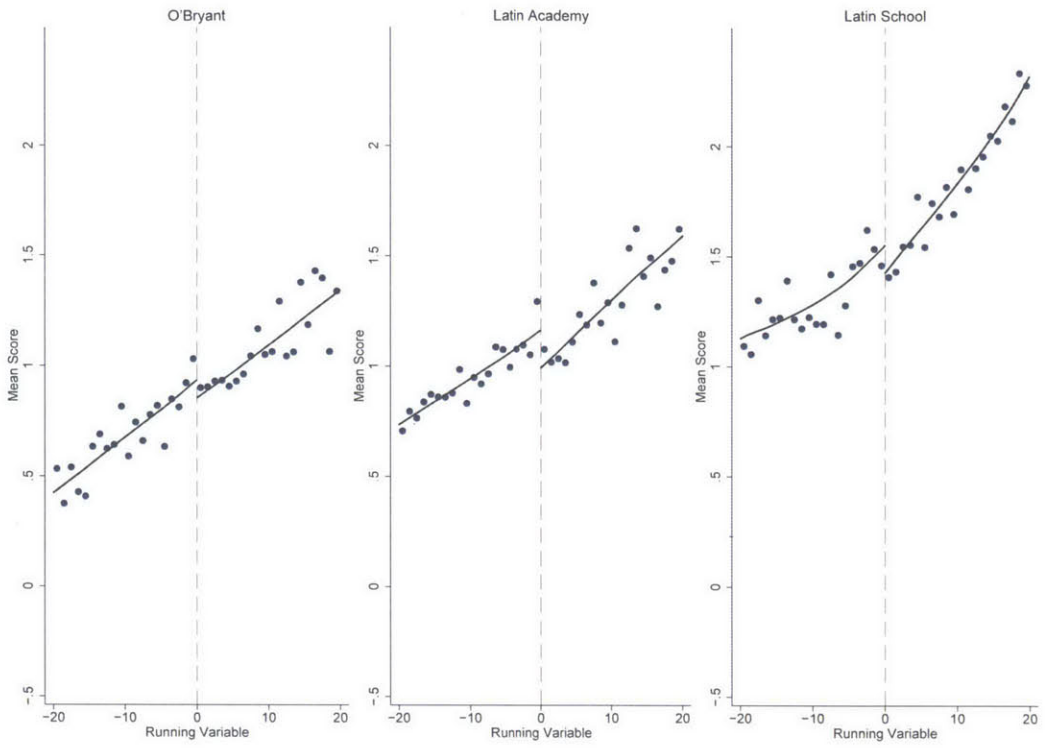


(a) Exam School Offer

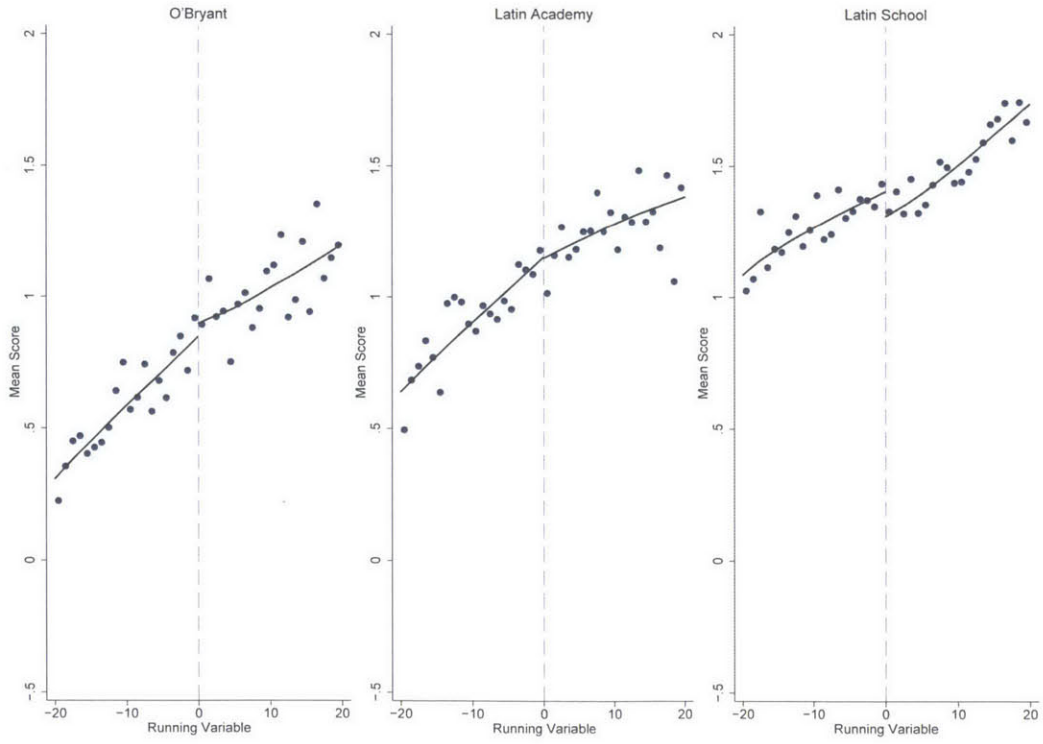


(b) Exam School Enrollment

Figure 1-6: Relationship between Exam School Offer and Enrollment and the Running Variables

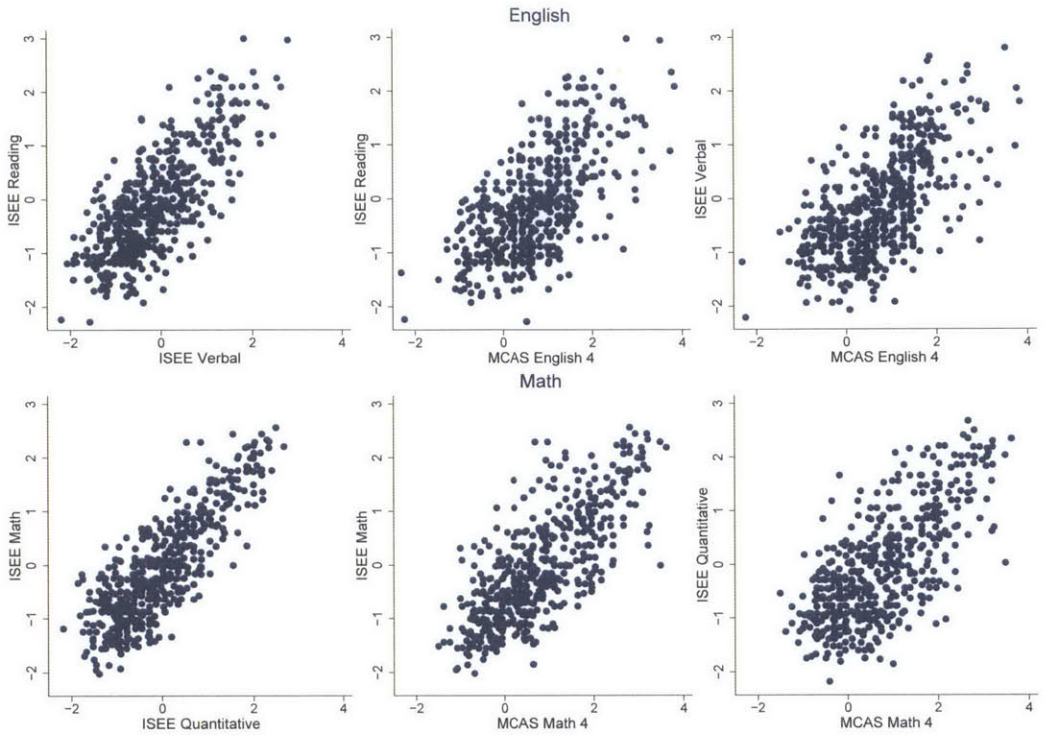


(a) Middle School MCAS Composite

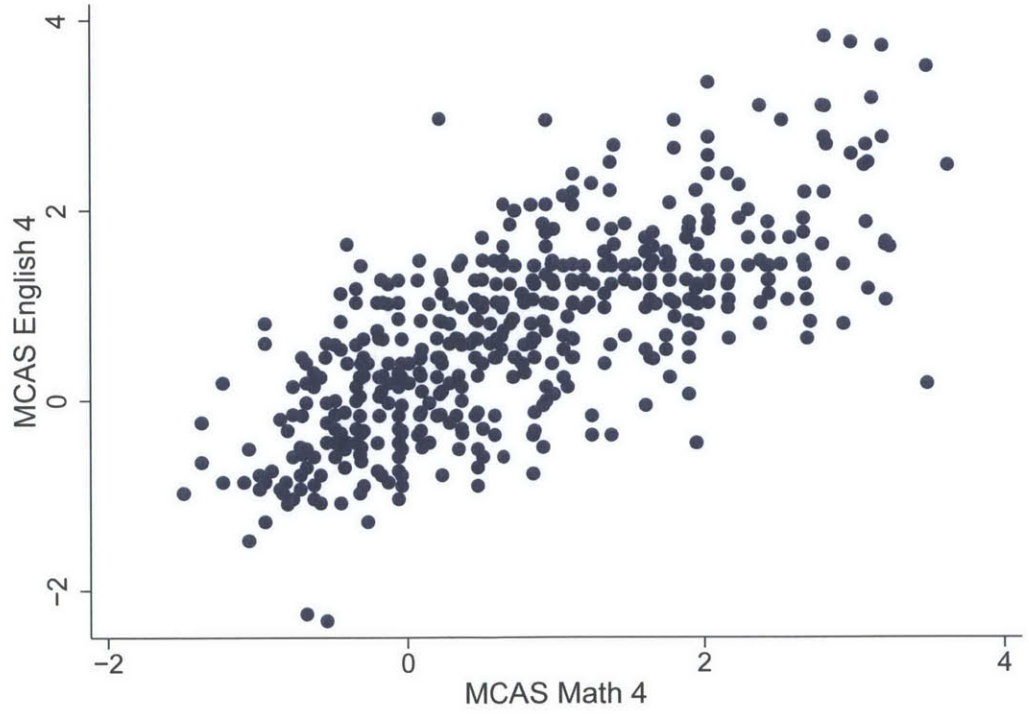


(b) High School MCAS Composite

Figure 1-7: Relationship between Middle School and High School MCAS Composites and the Running Variables

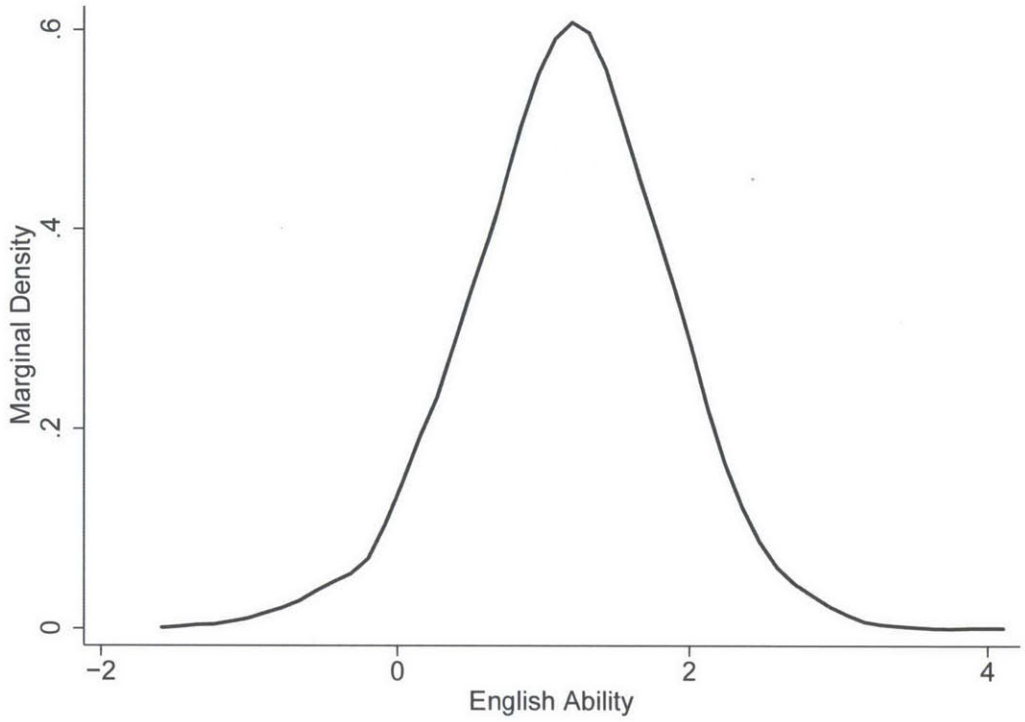


(a) ISEE Scores and 4th Grade MCAS Scores

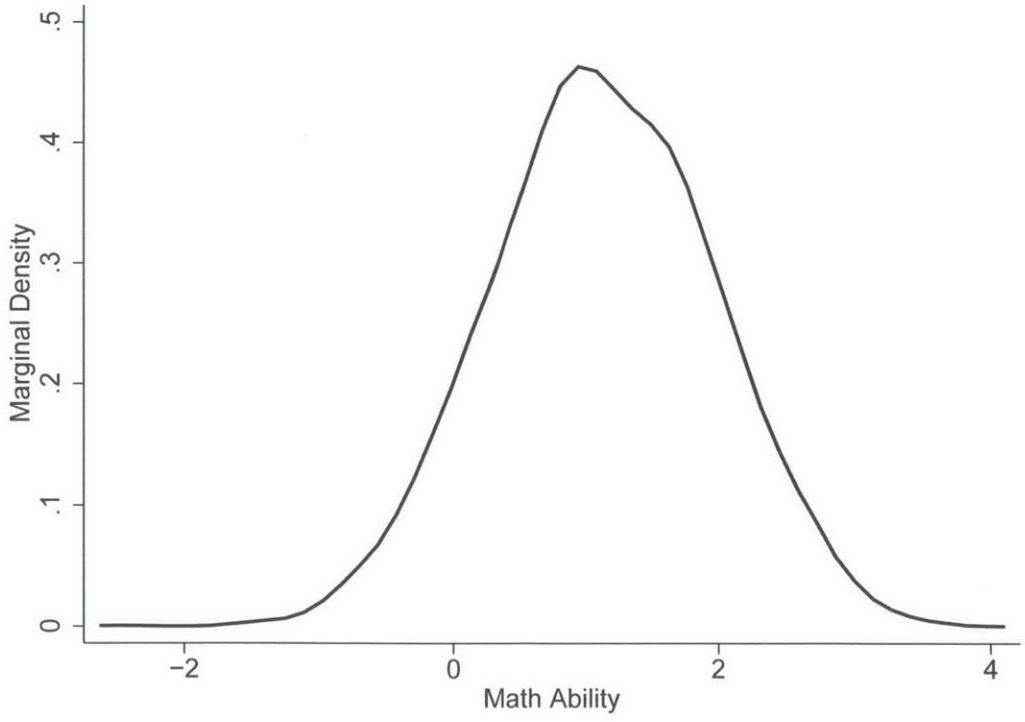


(b) 4th Grade MCAS Scores

Figure 1-8: Scatterplots of ISEE Scores and 4th Grade MCAS Scores



(a) Marginal Distribution of English Ability



(b) Marginal Distribution of Math Ability

Figure 1-9: Marginal Distributions of the English and Math Abilities

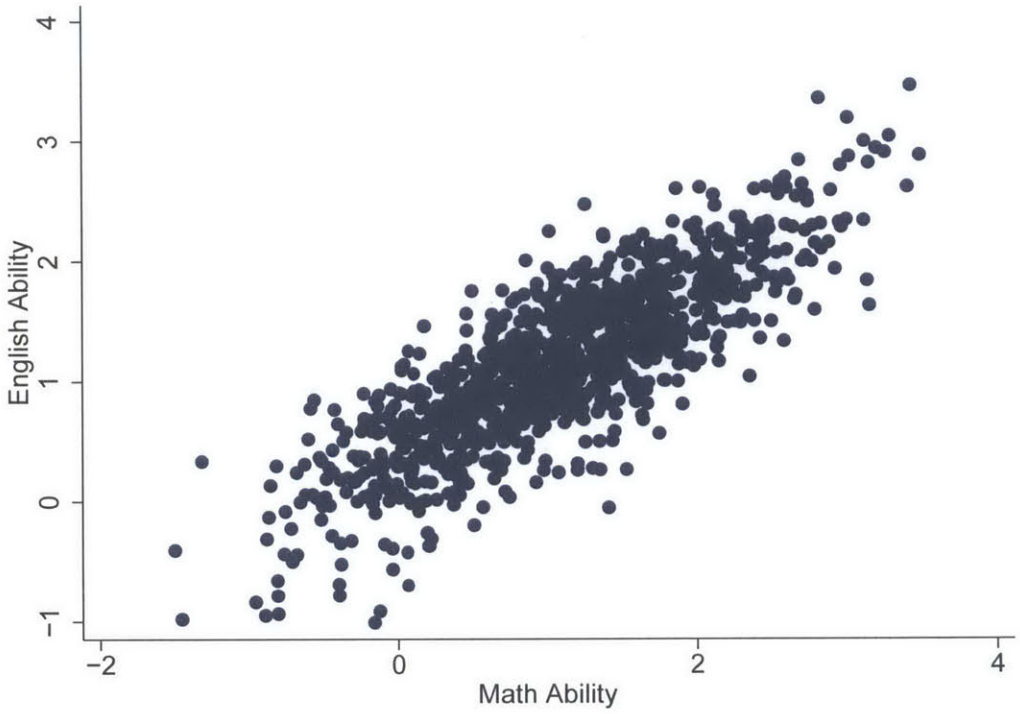
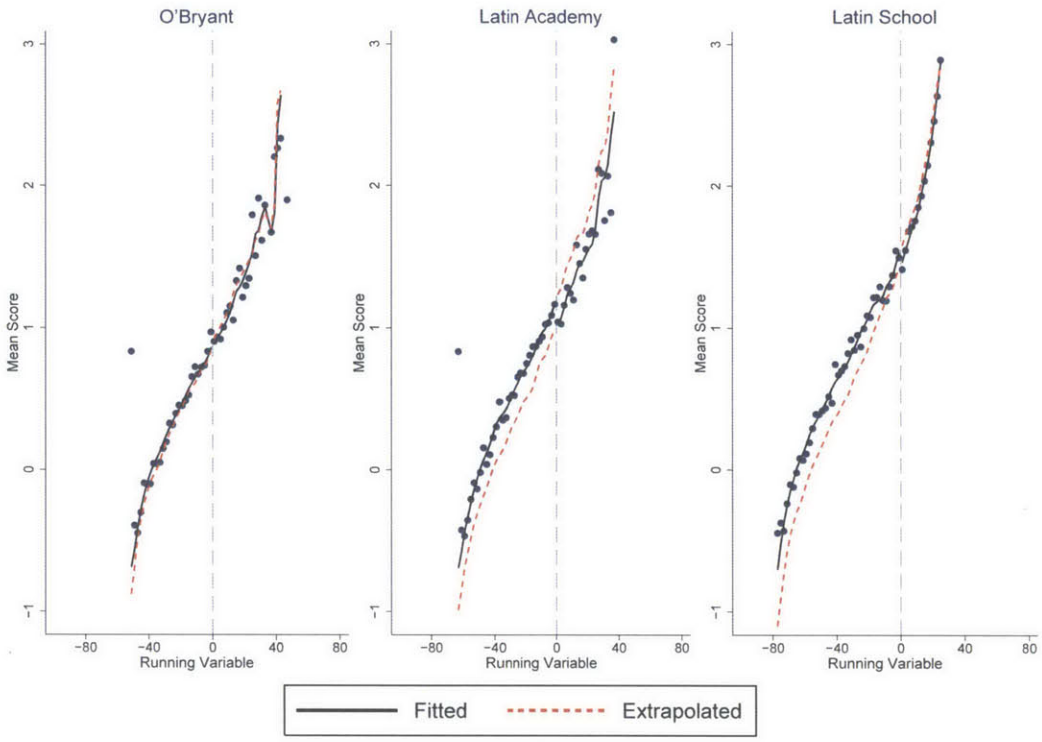
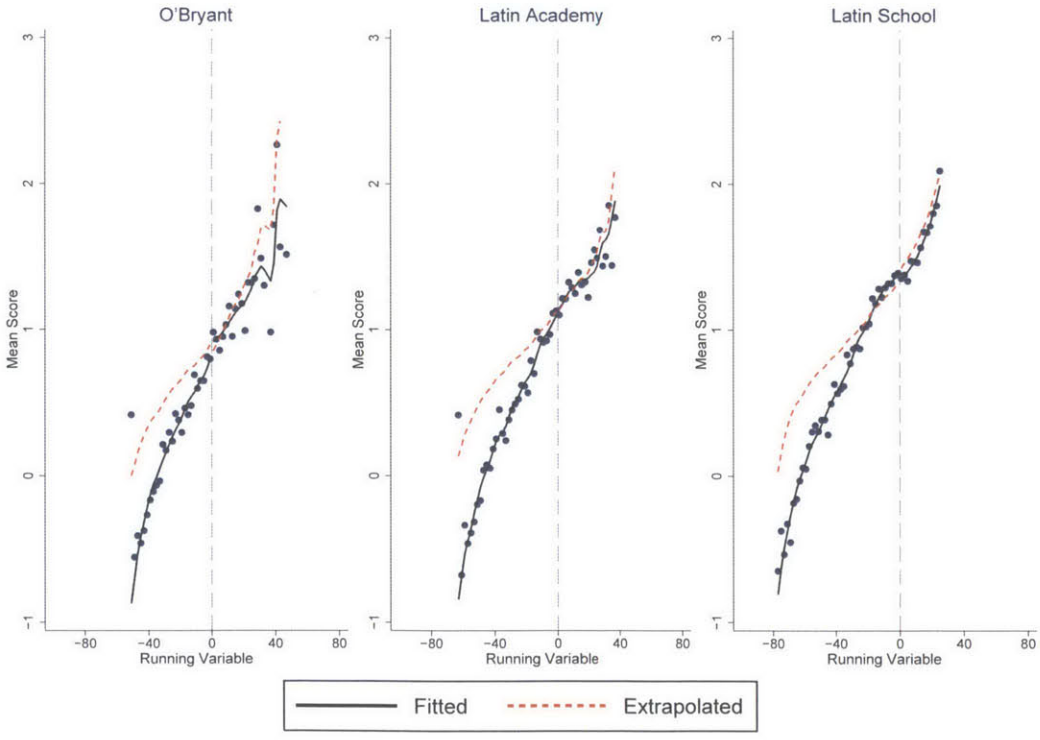


Figure 1-10: Scatterplot of English and Math Abilities

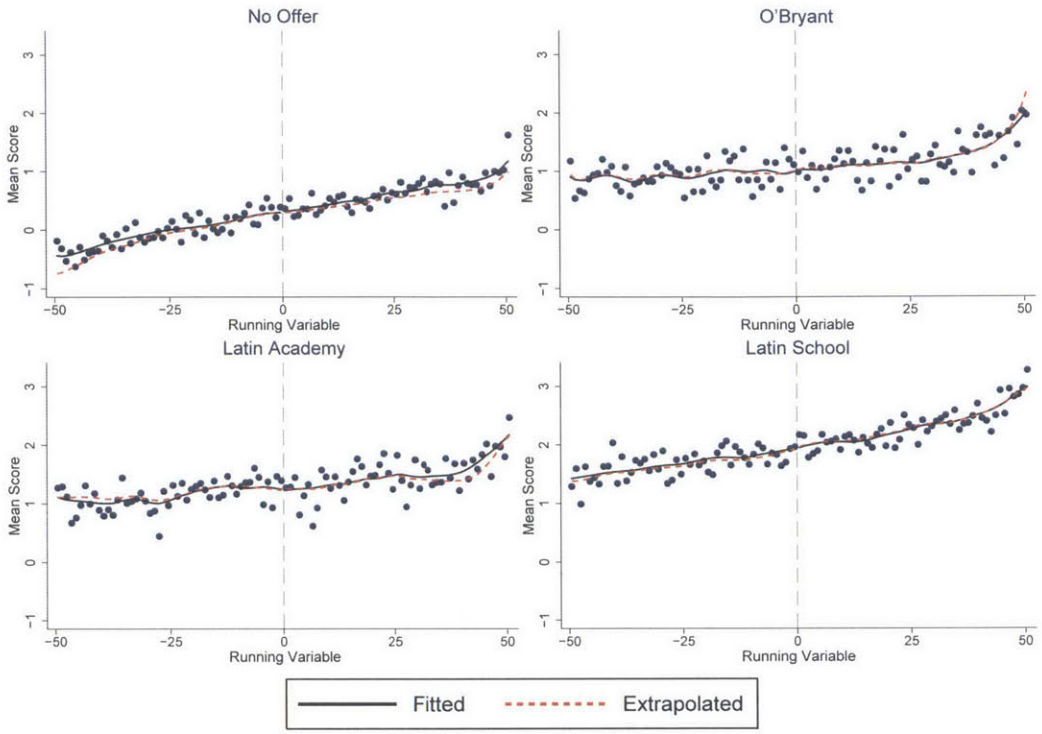


(a) Middle School MCAS Composite

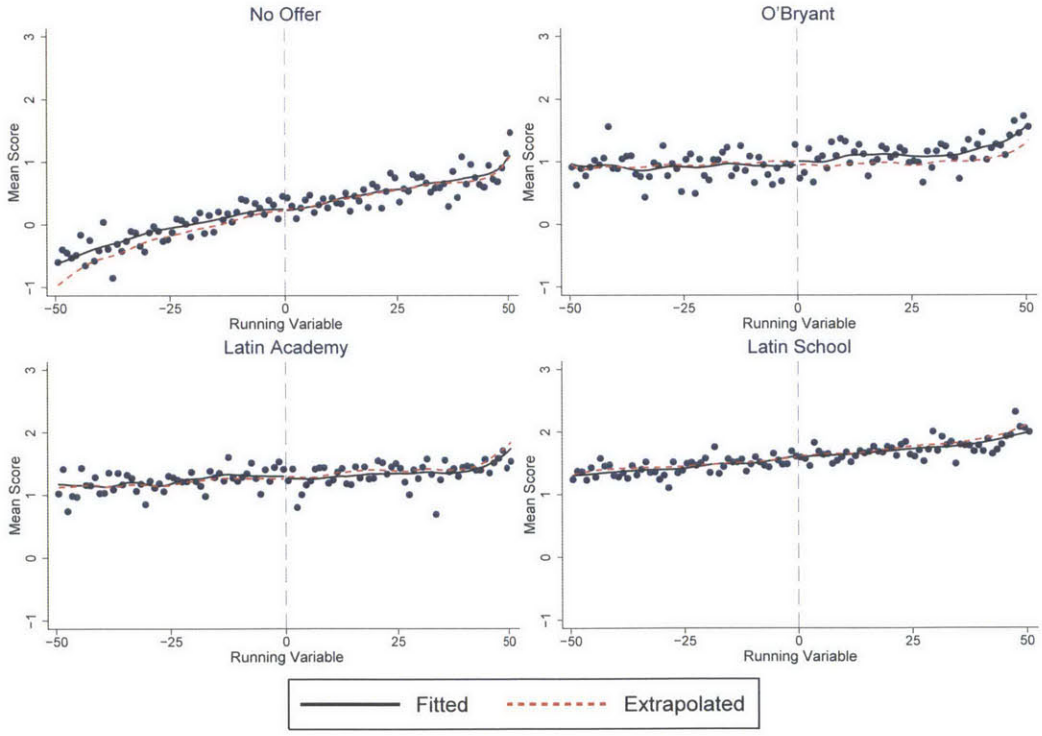


(b) Middle School MCAS Composite

Figure 1-11: Extrapolated Reduced Form Effects



(a) Middle School MCAS Composite



(b) High School MCAS Composite

Figure 1-12: Extrapolated Reduced Form Effects in the Placebo RD Experiments

Table 1.1: Descriptive Statistics for Boston Public School Students and Exam School Applicants

			Exam School Assignment			
	All BPS	All Applicants	No Offer	O'Bryant	Latin Academy	Latin School
	(1)	(2)	(3)	(4)	(5)	(6)
Female	0.489	0.545	0.516	0.579	0.581	0.577
Black	0.516	0.399	0.523	0.396	0.259	0.123
Hispanic	0.265	0.189	0.223	0.196	0.180	0.081
FRPL	0.755	0.749	0.822	0.788	0.716	0.499
LEP	0.116	0.073	0.109	0.064	0.033	0.004
Bilingual	0.315	0.387	0.353	0.420	0.451	0.412
SPED	0.227	0.043	0.073	0.009	0.006	0.009
English 4	0.000	0.749	0.251	0.870	1.212	1.858
Math 4	0.000	0.776	0.206	0.870	1.275	2.114
N	21,094	5,179	2,791	755	790	843

Notes: This table reports descriptive statistics for 2000-2004. The All BPS column includes all 6th grade students in Boston Public Schools in who do not have missing covariate or 4th grade MCAS information. The All Applicants column includes the subset of students who apply to Boston exam schools. The Assignment columns include the subsets of applicants who receive an offer from a given exam school.

Table 1.2: RD Estimates for the First Stage, Reduced Form and Local Average Treatment Effects at the Admissions Cutoffs

	O'Bryant	Latin Academy	Latin School
	(1)	(2)	(3)
	Panel A: Middle School MCAS		
First Stage	0.775***	0.949***	0.962***
	(0.031)	(0.017)	(0.017)
Reduced Form	-0.084	-0.181***	-0.104
	(0.060)	(0.057)	(0.079)
LATE	-0.108	-0.191***	-0.108
	(0.078)	(0.060)	(0.082)
N	1,934	2,328	1,008
	Panel B: High School MCAS		
First Stage	0.781***	0.955***	0.964***
	(0.034)	(0.018)	(0.018)
Reduced Form	0.047	-0.021	-0.086
	(0.055)	(0.044)	(0.052)
LATE	0.060	-0.022	-0.089
	(0.070)	(0.046)	(0.054)
N	1,475	1,999	907

Notes: This table reports RD estimates of the effect of an exam school offer on exam school enrollment (First Stage), the effect of an exam school offer on MCAS scores (Reduced Form), and the effects of exam school enrollment on MCAS scores (LATE). Heteroskedasticity-robust standard errors shown in parentheses.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 1.3: RD Estimates for the First Stage, Reduced Form and Local Average Treatment Effects at the Admissions Cutoffs: Heterogeneity by Average 4th Grade MCAS Scores

	Low 4th Grade MCAS Composite			High 4th Grade MCAS Composite		
	O'Bryant	Latin Academy	Latin School	O'Bryant	Latin Academy	Latin School
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Middle School MCAS					
First Stage	0.734***	0.917***	1.000***	0.780***	0.956***	0.961***
	(0.072)	(0.056)	(0.000)	(0.033)	(0.017)	(0.017)
Reduced Form	0.059	0.130	-0.095	-0.122*	-0.200***	-0.084
	(0.127)	(0.199)	(0.214)	(0.065)	(0.056)	(0.082)
LATE	0.080	0.142	-0.095	-0.157*	-0.209***	-0.088
	(0.173)	(0.215)	(0.214)	(0.083)	(0.059)	(0.085)
N	420	246	681	1,348	1,802	921
	Panel B: High School MCAS					
First Stage	0.742***	0.944***	1.000***	0.797***	0.942***	0.964***
	(0.074)	(0.055)	(0.000)	(0.033)	(0.022)	(0.017)
Reduced Form	0.204**	-0.029	-0.272	0.017	-0.029	-0.093*
	(0.099)	(0.141)	(0.186)	(0.054)	(0.052)	(0.053)
LATE	0.275**	-0.031	-0.272	0.021	-0.030	-0.097*
	(0.136)	(0.150)	(0.186)	(0.068)	(0.055)	(0.055)
N	531	567	436	1,129	974	909

Notes: This table reports RD estimates of the effect of an exam school offer on exam school enrollment (First Stage), the effect of an exam school offer on MCAS scores (Reduced Form), and the effects of exam school enrollment on MCAS scores (LATE). The estimates are shown separately for applicants whose average 4th grade MCAS scores fall below and above the within-year median. Heteroskedasticity-robust standard errors shown in parentheses.

* significant at 10%, ** significant at 5%, *** significant at 1%

Table 1.4: Correlations between the ISEE Scores and 4th Grade MCAS Scores

	ISEE				MCAS	
	Reading	Verbal	Math	Quantitative	English 4	Math 4
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: ISEE			
Reading	1	0.735	0.631	0.621	0.670	0.581
Verbal	0.735	1	0.619	0.617	0.655	0.587
Math	0.631	0.619	1	0.845	0.598	0.740
Quantitative	0.621	0.617	0.845	1	0.570	0.718
			Panel B: MCAS			
English 4	0.670	0.655	0.598	0.570	1	0.713
Math 4	0.581	0.587	0.740	0.718	0.713	1.000
N			5,179			

Notes: This table reports correlations between the ISEE scores and 4th grade MCAS scores.

Table 1.5: Factor Loadings on the Means and (Log) Standard Deviations of the ISEE and 4th Grade MCAS Scores

	ISEE				MCAS	
	Reading	Verbal	Math	Quantitative	English 4	Math 4
	(1)	(2)	(3)	(4)	(5)	(6)
			Panel A: Factor Loading on Mean			
θ_E	1.160***	1.180***			1	
	(0.029)	(0.032)				
θ_M			1.135***	1.119***		1
			(0.025)	(0.024)		
			Panel B: Factor Loading on (Log) Standard Deviation			
θ_E	0.081***	0.152***			0.016	
	(0.023)	(0.015)			(0.016)	
θ_M			0.019	-0.013		0.124***
			(0.020)	(0.016)		(0.014)
N			5,179			

Notes: This table reports the estimated factor loadings on the means and (log) standard deviations of the ISEE and 4th grade MCAS scores. Standard errors based on nonparametric 5-step boostrap shown in paretheses.

* significant at 10%, ** significant at 5%, *** significant at 1%

Table 1.6: Factor Loadings on Enrollment and MCAS Scores Under a Given Exam School Assignment

	No Offer	O'Bryant	Latin Academy	Latin School
	(1)	(2)	(3)	(4)
	Panel A: Middle School MCAS			
First Stage				
θ_E	-0.001	-0.073	-0.107**	0.016
	(0.007)	(0.069)	(0.050)	(0.011)
θ_M	-0.008	-0.035	0.011	0.023
	(0.006)	(0.059)	(0.031)	(0.015)
Reduced Form				
θ_E	0.570***	0.476***	0.190**	0.439***
	(0.052)	(0.087)	(0.089)	(0.063)
θ_M	0.633***	0.601***	0.742***	0.464***
	(0.050)	(0.067)	(0.067)	(0.063)
N	2,490	690	728	793
	Panel B: High School MCAS			
First Stage				
θ_E	-0.001	-0.025	-0.102*	0.003
	(0.011)	(0.076)	(0.053)	(0.009)
θ_M	-0.012	-0.059	0.012	0.027**
	(0.008)	(0.061)	(0.030)	(0.012)
Reduced Form				
θ_E	0.446***	0.282***	0.217***	0.239***
	(0.062)	(0.067)	(0.066)	(0.046)
θ_M	0.604***	0.346***	0.267***	0.158***
	(0.056)	(0.070)	(0.057)	(0.044)
N	1,777	563	625	793

Notes: This table reports the estimated factor loadings on enrollment (First Stage) and MCAS scores (Reduced Form) under a given exam school assignment. First Stage refers to enrollment at a traditional Boston public school in the No Offer column and enrollment at a given exam school in the other columns. Standard errors based on nonparametric 5-step boostrap shown in paretheses.

* significant at 10%, ** significant at 5%, *** significant at 1%

Table 1.7: Extrapolated First Stage, Reduced Form, and Local Average Treatment Effects in the Exam School-Specific RD Experiments

	O'Bryant	Latin Academy	Latin School
	(1)	(2)	(3)
	Panel A: Middle School MCAS		
First Stage	0.869***	0.971***	0.950***
	(0.033)	(0.013)	(0.017)
Reduced Form	-0.047	-0.229**	-0.214***
	(0.093)	(0.100)	(0.080)
LATE	-0.054	-0.236**	-0.226***
	(0.108)	(0.103)	(0.084)
N	3,029	3,641	4,271
	Panel B: High School MCAS		
First Stage	0.858***	0.962***	0.950***
	(0.038)	(0.021)	(0.018)
Reduced Form	0.252***	0.279***	0.199***
	(0.068)	(0.075)	(0.061)
LATE	0.293***	0.290***	0.209***
	(0.081)	(0.079)	(0.064)
N	2,240	2,760	3,340

Notes: This table reports latent factor model based-estimates of the effect of an exam school offer on exam school enrollment (First Stage), the effect of an exam school offer on MCAS scores (Reduced Form), and the effects of exam school enrollment on MCAS scores (LATE) in the RD experiments. Standard errors based on nonparametric 5-step boostrap shown in paretheses.
* significant at 10%, ** significant at 5%, *** significant at 1%

Table 1.8: Extrapolated First Stage, Reduced Form, and Local Average Treatment Effects in the Exam School-Specific RD Experiments: Heterogeneity by the Running Variables

	Below Admissions Cutoff			Above Admissions Cutoff		
	O'Bryant	Latin Academy	Latin School	O'Bryant	Latin Academy	Latin School
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Middle School MCAS					
First Stage	0.904***	0.992***	0.960***	0.749***	0.891***	0.906***
	(0.041)	(0.015)	(0.020)	(0.016)	(0.018)	(0.031)
Reduced Form	-0.050	-0.234*	-0.245**	-0.038	-0.211***	-0.080
	(0.118)	(0.123)	(0.097)	(0.034)	(0.061)	(0.076)
LATE	-0.055	-0.236*	-0.255**	-0.051	-0.237***	-0.088
	(0.132)	(0.124)	(0.101)	(0.045)	(0.069)	(0.085)
N	2,339	2,913	3,478	690	728	793
	Panel B: High School MCAS					
First Stage	0.892***	0.984***	0.959***	0.758***	0.885***	0.920***
	(0.050)	(0.027)	(0.021)	(0.017)	(0.021)	(0.029)
Reduced Form	0.343***	0.362***	0.281***	-0.021	-0.005	-0.091*
	(0.089)	(0.092)	(0.075)	(0.034)	(0.046)	(0.054)
LATE	0.385***	0.367***	0.293***	-0.028	-0.006	-0.099*
	(0.102)	(0.097)	(0.079)	(0.045)	(0.053)	(0.059)
N	1,677	2,135	2,601	563	625	739

Notes: This table reports latent factor model based-estimates of the effect of an exam school offer on exam school enrollment (First Stage), the effect of an exam school offer on MCAS scores (Reduced Form), and the effects of exam school enrollment on MCAS scores (LATE) in the RD experiments. The estimates are shown separately for applicants whose running variables fall below and above the admissions cutoffs. Standard errors based on nonparametric 5-step boostrap shown in paretheses.

* significant at 10%, ** significant at 5%, *** significant at 1%

Table 1.9: Extrapolated First Stage, Reduced Form, and Local Average Treatment Effects for Comparisons between a Given Exam School and Traditional Boston Public Schools

	O'Bryant	Latin Academy	Latin School
	(1)	(2)	(3)
	Panel A: Middle School MCAS		
First Stage	0.767***	0.956***	0.967***
	(0.017)	(0.009)	(0.016)
Reduced Form	-0.059	-0.275***	-0.319***
	(0.048)	(0.074)	(0.079)
LATE	-0.077	-0.288***	-0.330***
	(0.063)	(0.078)	(0.082)
N		4,701	
	Panel B: High School MCAS		
First Stage	0.754***	0.948***	0.968***
	(0.020)	(0.013)	(0.016)
Reduced Form	0.021	0.049	0.024
	(0.037)	(0.058)	(0.064)
LATE	0.027	0.052	0.025
	(0.049)	(0.062)	(0.066)
N		3,704	

Notes: This table reports latent factor model based-estimates of the effect of receiving an offer from a given exam school versus no offer at all on enrollment at this exam school (First Stage), the effect of receiving an offer from a given exam school versus no offer at all on MCAS scores (Reduced Form), and the effect of enrollment at this exam school versus a traditional Boston public school on MCAS scores (LATE). Standard errors based on nonparametric 5-step boostrap shown in paretheses.

* significant at 10%, ** significant at 5%, *** significant at 1%

Table 1.10: Extrapolated First Stage, Reduced Form, and Local Average Treatment Effects for Comparisons between a Given Exam School and Traditional Boston Public Schools: Heterogeneity by Exam School Offer Status

	No Exam School Offer			Exam School Offer		
	O'Bryant	Latin Academy	Latin School	O'Bryant	Latin Academy	Latin School
	(1)	(2)	(3)	(4)	(5)	(6)
	Panel A: Middle School MCAS					
First Stage	0.901***	0.994***	0.958***	0.749***	0.927***	0.986***
	(0.040)	(0.016)	(0.024)	(0.016)	(0.010)	(0.005)
Reduced Form	-0.051	-0.246*	-0.279**	-0.069	-0.307***	-0.363***
	(0.117)	(0.138)	(0.117)	(0.074)	(0.048)	(0.051)
LATE	-0.057	-0.248*	-0.292**	-0.092	-0.332***	-0.368***
	(0.131)	(0.140)	(0.122)	(0.099)	(0.052)	(0.052)
N		2,490			2,211	
	Panel B: High School MCAS					
First Stage	0.887***	0.987***	0.955***	0.758***	0.925***	0.987***
	(0.049)	(0.030)	(0.026)	(0.017)	(0.010)	(0.005)
Reduced Form	0.334***	0.429***	0.428***	-0.268***	-0.300***	-0.348***
	(0.088)	(0.105)	(0.094)	(0.066)	(0.048)	(0.052)
LATE	0.376***	0.435***	0.448***	-0.353***	-0.325***	-0.353***
	(0.102)	(0.110)	(0.098)	(0.087)	(0.053)	(0.053)
N		1,777			1,927	

Notes: This table reports latent factor model based-estimates of the effect of receiving an offer from a given exam school versus no offer at all on enrollment at this exam school (First Stage), the effect of receiving an offer from a given exam school versus no offer at all on MCAS scores (Reduced Form), and the effect of enrollment at this exam school versus a traditional Boston public school on MCAS scores (LATE). The estimates are shown separately for applicants who do not receive an exam school offer and for applicants who receive an exam school offer. Standard errors based on nonparametric 5-step boostrap shown in paretheses.

* significant at 10%, ** significant at 5%, *** significant at 1%

Table 1.11: Extrapolated Reduced Form Effects in Placebo RD Experiments

	No Offer (1)	O'Bryant (2)	Latin Academy (3)	Latin School (4)
		Panel A: All Applicants		
Middle	0.002	-0.000	0.042	-0.023
School	(0.063)	(0.089)	(0.071)	(0.072)
MCAS	2,490	690	728	793
High	-0.057	0.080	-0.041	-0.008
School	(0.067)	(0.071)	(0.058)	(0.051)
MCAS	1,777	563	625	793
		Panel B: Below Placebo Cutoff		
Middle	-0.081	0.015	0.034	-0.034
School	(0.099)	(0.070)	(0.057)	(0.065)
MCAS	1,244	344	362	395
High	-0.133	0.027	-0.028	0.027
School	(0.105)	(0.064)	(0.057)	(0.049)
MCAS	887	280	311	368
		Panel C: Above Placebo Cutoff		
Middle	0.085	-0.015	0.050	-0.013
School	(0.053)	(0.131)	(0.125)	(0.112)
MCAS	1,246	346	366	395
High	0.020	0.133	-0.054	-0.044
School	(0.059)	(0.099)	(0.091)	(0.082)
MCAS	890	283	314	371

Notes: This table reports latent factor model-based estimates of the effects of placebo offers on MCAS scores. The estimates are shown for all applicants and separately for applicants whose running variables fall below and above the placebo admissions cutoffs. Standard errors based on nonparametric 5-step boostrap shown in paretheses.
* significant at 10%, ** significant at 5%, *** significant at 1%

Table 1.12: Actual and Counterfactual Assignments under Minority and Socioeconomic Preferences

	Actual Assignment			
Counterfactual Assignment	No Offer (1)	O'Bryant (2)	Latin Academy (3)	Latin School (4)
	Panel A: Minority Preferences			
No Offer	2418	221	113	39
O'Bryant	280	129	133	213
Latin Academy	88	389	268	45
Latin School	5	16	276	546
	Panel B: Socioeconomic Preferences			
No Offer	2579	159	39	14
O'Bryant	203	319	146	87
Latin Academy	9	106	403	272
Latin School	0	171	202	470

Notes: This table reports the actual assignments and the counterfactual assignments under minority and socioeconomic preferences in the exam school admissions.

Table 1.13: Counterfactual Admissions Cutoffs for Different Applicant Groups under Minority and Socioeconomic Preferences

	O'Bryant (1)	Latin Academy (2)	Latin School (3)
	Panel A: Minority Preferences		
Minority	-14.1	-20.6	-31.9
Non-Minority	15.8	12.4	7.8
	Panel B: Socioeconomic Preferences		
SES Tier 1	-20.4	-26.4	-32.9
SES Tier 2	-6.6	-11.7	-16.1
SES Tier 3	-2.5	-7.2	-17.1
SES Tier 4	8.0	2.1	-5.1

Notes: This table reports the differences between the actual admissions cutoffs and the counterfactual admissions cutoffs under minority and socioeconomic preferences in the exam school admissions.

Table 1.14: Composition of Applicants by the Counterfactual Assignment under Minority and Socioeconomic Preferences

	No Offer (1)	O'Bryant (2)	Latin Academy (3)	Latin School (4)
		Panel A: Minority Preferences		
Female	0.502	0.588	0.629	0.572
Black	0.430	0.440	0.386	0.274
Hispanic	0.182	0.220	0.203	0.172
FRPL	0.810	0.771	0.715	0.556
LEP	0.116	0.030	0.022	0.018
Bilingual	0.386	0.396	0.380	0.391
SPED	0.073	0.009	0.013	0.005
English 4	0.277	0.981	1.215	1.668
Math 4	0.277	0.963	1.289	1.781
		Panel B: Minority Preferences		
Female	0.514	0.572	0.597	0.575
Black	0.499	0.360	0.295	0.203
Hispanic	0.223	0.191	0.143	0.120
FRPL	0.813	0.728	0.673	0.624
LEP	0.107	0.058	0.030	0.017
Bilingual	0.359	0.423	0.391	0.445
SPED	0.073	0.012	0.006	0.006
English 4	0.261	1.067	1.309	1.556
Math 4	0.217	1.146	1.365	1.744
N	2,791	755	790	843

Notes: This table reports descriptive statistics for the exam school applicants by their counterfactual assignment under minority and socioeconomic preferences in the exam school admissions.

Table 1.15: Average Reassignment Effects of Introducing Minority or Socioeconomic Preferences into the Boston Exam School Admissions

	Minority Preferences			Socioeconomic Preferences				
		Applicant Group			Applicant Group			
	All Applicants	Minority	Non-Minority	All Applicants	SES Tier 1	SES Tier 2	SES Tier 3	SES Tier 4
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Panel A: All Applicants				
Middle	0.001	-0.028***	0.043***	0.007	-0.032***	-0.005	0.004	0.041***
School	(0.005)	(0.007)	(0.010)	(0.005)	(0.011)	(0.006)	(0.007)	(0.011)
MCAS	4,701	2,741	1,960	4,701	948	1,113	1,155	1,468
High	0.023***	0.027***	0.017*	0.015***	0.024**	0.011*	0.006	0.018*
School	(0.004)	(0.007)	(0.009)	(0.005)	(0.011)	(0.006)	(0.005)	(0.010)
MCAS	3,704	2,086	1,618	3,704	737	876	901	1,190
				Panel B: Affected Applicants				
Middle	0.003	-0.084***	0.113***	0.024	-0.092***	-0.021	0.021	0.133***
School	(0.014)	(0.021)	(0.026)	(0.019)	(0.032)	(0.027)	(0.031)	(0.034)
MCAS	1,670	932	738	1,294	328	263	241	426
High	0.062***	(0.025)	0.046**	0.050***	0.068**	0.044*	0.026	0.054*
School	(0.011)	(0.017)	(0.020)	(0.016)	(0.031)	(0.023)	(0.024)	(0.029)
MCAS	1,367	754	613	1,100	265	223	208	404

Notes: This table reports the latent factor model-based estimates of the effects of minority and socioeconomic preferences on MCAS scores. The estimates are shown for all applicants and separately for the applicant groups who face different admissions cutoffs after the reforms. The estimates are also shown separately for the affected applicants whose exam school assignment is altered by the reforms. Standard errors based on nonparametric 5-step boostrap shown in paretheses.

* significant at 10%, ** significant at 5%, *** significant at 1%

1.9 Appendix A: Proofs

Proof of Lemma 2 The result follows directly from the Law of Iterated Expectations. ■

Proof of Theorem 1 Notice first that the means of M_1 and M_2 as well as covariances between M_1 and M_3 and between M_2 and M_3 can be written as

$$\begin{aligned} E[M_1] &= E[\theta] \\ E[M_2] &= \mu_{M_2} + \lambda_{M_2} E[\theta] \\ Cov[M_1, M_3] &= Cov[\theta, W] \\ Cov[M_2, M_3] &= \lambda_{M_2} Cov[\theta, M_3]. \end{aligned}$$

From these equations one can solve for the parameters μ_{M_2} and λ_{M_2} that are given by

$$\begin{aligned} \lambda_{M_2} &= \frac{Cov[M_2, M_3]}{Cov[\theta, M_3]} \\ \mu_{M_2} &= E[M_2] - \lambda_{M_2} E[M_1]. \end{aligned}$$

Let us now introduce two new random variables, $\tilde{M}_2$ and $\tilde{\nu}_{M_2}$, that are defined as

$$\begin{aligned} \tilde{M}_2 &= \frac{1}{\lambda_{M_2}} (M_2 - \mu_{M_2}) \\ \tilde{\nu}_{M_2} &= \frac{1}{\lambda_M} \nu_{M_2}. \end{aligned}$$

Thus, M_1 and $\tilde{M}_2$ can be written as

$$\begin{aligned} M_1 &= \theta + \nu_{M_1} \\ \tilde{M}_2 &= 0 + \tilde{\nu}_{M_2}. \end{aligned}$$

Depending on whether Assumption C.4.a or Assumption C.4.b holds, M_1 and $\tilde{M}_2$ satisfy either Assumption A or Assumption B in Evdokimov and White (2012) conditional on M_3. In their notation $M = \theta$ and, depending on the assumptions imposed on ν_{M_1} and ν_{M_2}, either $Y_1 = M_1$, $Y_2 = \tilde{M}_2$, $U_1 = \nu_{M_1}$, and $U_2 = \tilde{\nu}_{M_2}$ or $Y_1 = \tilde{M}_2$, $Y_2 = M_1$, $U_1 = \tilde{\nu}_{M_2}$, and $U_2 = \nu_{M_1}$. The identification of the conditional distributions $f_{\nu_{M_1}|M_3}$, $f_{\tilde{\nu}_{M_2}|M_3}$ and $f_{\theta|M_3}$ follows from either Lemma 1 or Lemma 2 in Evdokimov and White (2012) depending on whether Assumption A or Assumption B is satisfied. The conditional distribution $f_{\nu_{M_2}|M_3}$ and the joint

distribution $f_{\theta,M}$ are given by

$$\begin{aligned} f_{\nu_{M_2}|M_3}(\nu_{M_2} \mid m_3) &= \frac{1}{\lambda_{M_2}} f_{\tilde{\nu}_{M_2}|M_3}\left(\frac{1}{\lambda_{M_2}}\nu_{M_2} \mid m_3\right) \\ f_{\theta,M}(\theta, m) &= f_{\nu_{M_1}|M_3}(m_1 - \theta \mid m_3) f_{\nu_{M_2}|M_3}(m_2 - \mu_{M_2} - \lambda_{M_2}\theta \mid m_3) \\ &\quad \times f_{\theta|M_3}(\theta \mid m_3) f_{M_3}(m_3). \end{aligned}$$

■

Proof of Theorem 2 Assumptions D.1, D.3, D.4, and D.5 correspond to assumptions 1, 3, 4, and 5 in Hu and Schennach (2008) with $y = M_3$, $x = M_1$, $z = M_2$, and $x^* = \theta$ in their notation. Furthermore, as shown in Cunha, Heckman, and Schennach (2010), Assumption D.2 is equivalent to Assumption 2 in Hu and Schennach (2008). The identification of the conditional distributions $f_{M_1|\theta}$, $f_{M_3|\theta}$, and $f_{\theta|M_2}$ then follows from Theorem 1 in Hu and Schennach (2008). The joint distribution $f_{\theta,M}$ is given by

$$f_{\theta,M}(\theta, m) = f_{M_1|\theta}(m_1 \mid \theta) f_{M_3|\theta}(m_3 \mid \theta) f_{\theta|M_2}(\theta \mid m_2) f_{M_2}(m_2).$$

■

Proof of Theorem 3 Assumption E.1 allows one to write down the integral equations

$$\begin{aligned} E[Y \mid M = m^0, D = 0] &= E\{E[Y(0) \mid \theta] \mid M = m^0, D = 0\} \\ E[Y \mid M = m^1, D = 1] &= E\{E[Y(1) \mid \theta] \mid M = m^1, D = 1\}. \end{aligned}$$

The uniqueness of the solutions to these equations follows directly from Assumption E.3. To see this, suppose that in addition to $E[Y(0) \mid \theta]$ there exists some $\tilde{E}[Y(0) \mid \theta]$ such that

$$P\left\{E[Y(0) \mid \theta] \neq \tilde{E}[Y(0) \mid \theta]\right\} > 0$$

also satisfying the above equation for all $m^0 \in M^0$. Thus,

$$E\left\{E[Y(0) \mid \theta] - \tilde{E}[Y(0) \mid \theta] \mid R = r^0, D = 0\right\} = 0$$

for all $m^0 \in M^0$, and by Assumption E.3., this implies that $E[Y(0) \mid \theta] - \tilde{E}[Y(0) \mid \theta] = 0$ for all $m^0 \in M^0$, thus leading to a contradiction. An analogous argument can be given for the uniqueness of $E[Y(1) \mid \theta]$. Finally, Assumption E.2 guarantees that $E[Y(0) \mid \theta]$ and $E[Y(1) \mid \theta]$ are determined for all $\theta \in \Theta$. ■

Proof of Lemma 4 Using Assumptions H.1 and H.2, one can write

$$\begin{aligned}
&E\left\{E\left[Y\left(D\left(1\right)\right)-Y\left(D\left(0\right)\right)\mid\theta\right]\mid R=r\right\}\\
&=\int E\left[Y\left(D\left(1\right)\right)-Y\left(D\left(0\right)\right)\mid\theta\right]f_{\theta|R}\left(\theta\mid r\right)d\theta\\
&=\int E\left[Y\left(1\right)-Y\left(0\right)\mid D\left(1\right)>D\left(0\right),\theta\right]P\left[D\left(1\right)>D\left(0\right)\mid\theta\right]f_{\theta|R}\left(\theta\mid r\right)d\theta\\
&=\int E\left[Y\left(1\right)-Y\left(0\right)\mid D\left(1\right)>D\left(0\right),\theta,R=r\right]\\
&\quad\times P\left[D\left(1\right)>D\left(0\right)\mid\theta,R=r\right]f_{\theta|R}\left(\theta\mid r\right)d\theta. \qquad (1.5)
\end{aligned}$$

Furthermore, using the fact that

$$\begin{aligned}
&P\left[D\left(1\right)>D\left(0\right)\mid\theta,R=r\right]f_{\theta|R}\left(\theta\mid r\right)\\
&=f_{\theta|R}\left(\theta,D\left(1\right)>D\left(0\right)\mid r\right)\\
&=P\left[D\left(1\right)>D\left(0\right)\mid R=r\right]f_{\theta|R,D(0),D(1)}\left(\theta\mid r,0,1\right),
\end{aligned}$$

Equation (1.5) becomes

$$\begin{aligned}
&E\left\{E\left[Y\left(D\left(1\right)\right)-Y\left(D\left(0\right)\right)\mid\theta\right]\mid R=r\right\}\\
&=P\left[D\left(1\right)>D\left(0\right)\mid R=r\right]\\
&\quad\times\int E\left[Y\left(1\right)-Y\left(0\right)\mid D\left(1\right)>D\left(0\right),\theta,R=r\right]f_{\theta|R,D(0),D(1)}\left(\theta\mid r,0,1\right)d\theta\\
&=P\left[D\left(1\right)>D\left(0\right)\mid R=r\right]E\left[Y\left(1\right)-Y\left(0\right)\mid D\left(1\right)>D\left(0\right),R=r\right]. \qquad (1.6)
\end{aligned}$$

Using similar arguments, one can write

$$\begin{aligned}
&E\left\{E\left[D\left(1\right)-D\left(0\right)\mid\theta\right]\mid R=r\right\}\\
&=\int E\left[D\left(1\right)-D\left(0\right)\mid\theta\right]f_{\theta|R}\left(\theta\mid r\right)d\theta\\
&=\int P\left[D\left(1\right)>D\left(0\right)\mid\theta\right]f_{\theta|R}\left(\theta\mid r\right)d\theta\\
&=\int P\left[D\left(1\right)>D\left(0\right)\mid\theta,R=r\right]f_{\theta|R}\left(\theta\mid r\right)d\theta\\
&=P\left[D\left(1\right)>D\left(0\right)\mid R=r\right]\int f_{\theta|R,D(0),D(1)}\left(\theta\mid r,0,1\right)d\theta\\
&=P\left[D\left(1\right)>D\left(0\right)\mid R=r\right]. \qquad (1.7)
\end{aligned}$$

The result then follows from Equations (1.6) and (1.7). ■

Proof of Theorem 4 The proof is analgous to the proof of Theorem 3. ■

Proof of Theorem 5 Notice first that the identification of $\mu_{M_2^k}$, $\lambda_{M_2^k}$, $f_{\nu_{M_1^k|M_3}}$, $f_{\nu_{M_2^k}|M_3}$, and $f_{\theta_k|M_3}$ follows directly from Assumptions I.1-I.4 using Theorem 1 for all $k = 1, \ldots, K$. The only remaining issue is the identification of $f_{\theta|M_3}$. Let us start by defining new random variables

$$\begin{aligned}
\tilde{M}_2^k &= \frac{1}{\lambda_{M_2^k}} \left(M_2^k - \mu_{M_2^k} \right), k = 1, \ldots, K \\
\tilde{\nu}_{M_2^k} &= \frac{1}{\lambda_{M_2^k}} \nu_{M_2^k}, k = 1, \ldots, K.
\end{aligned}$$

Thus, $M_1 = \left(M_1^1, \ldots, M_1^K \right)$ and $\tilde{M}_2 = \left(\tilde{M}_2^1, \ldots, \tilde{M}_2^K \right)$ can be written as

$$\begin{aligned}
M_1 &= \theta + \nu_{M_1} \\
\tilde{M}_2 &= \theta + \tilde{\nu}_{M_2}
\end{aligned}$$

where $\theta = (\theta_1, \ldots, \theta_2)$ $\nu_{M_1} = \left(\nu_{M_1}^1, \ldots, \nu_{M_1}^K \right)$, and $\nu_{M_2} = \left(\nu_{M_2}^1, \ldots, \nu_{M_2}^K \right)$.

Now, using Assumption I.5, the conditional characteristic functions of M_1 and $\tilde{M}_2$ given M_3, $\phi_{M_1|M_3}$ and $\phi_{\tilde{M}_2|M_3}$, can be written as

$$\phi_{M_1|M_3}(m_1 \mid m_3) = \phi_{\theta|M_3}(m_1 \mid m_3) \phi_{\nu_{M_1}|M_3}(m_1 \mid m_3) \tag{1.8}$$

$$\phi_{\tilde{M}_2|M_3}(m_2 \mid m_3) = \phi_{\theta|M_3}(m_2 \mid m_3) \phi_{\tilde{\nu}_{M_2}|M_3}(m_2 \mid m_3) \tag{1.9}$$

where

$$\begin{aligned}
\phi_{\nu_{M_1}|M_3}(m_1 \mid m_3) &= \prod_{k=1}^{K} \phi_{\nu_{M_1^k}|M_3}\left(m_1^k \mid m_3 \right) \\
\phi_{\tilde{\nu}_{M_2}|M_3}(m_2 \mid m_3) &= \prod_{k=1}^{K} \phi_{\tilde{\nu}_{M_2}|M_3}\left(m_2^k \mid m_3 \right)
\end{aligned}$$

where the conditional characteristics functions $\phi_{\nu_{M_1^k}|M_3}$, $\phi_B(b)$, $\phi_{\tilde{\nu}_{M_2}|M_3}$, and $\phi_{\nu_B}(b)$ are known. Thus, the only unknown in this relationship is the conditional characteristic function of θ given M_3, $\phi_{\theta|M_3}$, that can be obtained from Equations (1.8) and (1.9):

$$\phi_{\theta|M_3}(t \mid m_3) = \begin{cases} \frac{\phi_{M_1|M_3}(t|m_3)}{\phi_{\tilde{\nu}_{M_2}|M_3}(t|m_3)}, \phi_{\tilde{\nu}_{M_2}|M_3}(t \mid m_3) \neq 0 \\ \frac{\phi_{\tilde{M}_2|M_3}(t|m_3)}{\phi_{\tilde{\nu}_{M_2}|M_3}(t|m_3)}, \phi_{\tilde{\nu}_{M_2}|M_3}(t \mid m_3) \neq 0 \end{cases}$$

Assumption I.6 quarantees that $\phi_{\theta|M_3}(t \mid m_3)$ is identified for all $t \in \mathbb{R}$.

Finally, the conditional characteristic functions $\phi_{\theta|M_3}$, $\phi_{\nu_{M_1}|M_3}$, and $\phi_{\tilde{\nu}_{M_2}|M_3}$uniquely determine the corresponding conditional distributions $f_{\theta|M_3}$, $f_{\nu_{M_1}|M_3}$, and $f_{\tilde{\nu}_{M_2}|M_3}$. The conditional distribution $f_{\nu_{M_2}|M_3}$

and the joint distribution $f_{\theta,M}$ are then given by

$$\begin{aligned} f_{\nu_{M_2}|M_3}(\nu_{M_2} \mid m_3) &= \prod_{k=1}^{K} \frac{1}{\lambda_{M_2}^k} f_{\tilde{\nu}_{M_2}^k|M_3}\left(\frac{1}{\lambda_{M_2}^k}\tilde{\nu}_{M_2}^k \mid m_3\right) \\ f_{\theta,M}(\theta, m) &= f_{\nu_{M_1}|M_3}(m_1 - \theta \mid m_3) f_{\nu_{M_2}|M_3}\left(m_2 - \mu_{M_2}^{'} - \lambda_{M_2}\theta^{'} \mid m_3\right) \\ &\quad \times f_{\theta|M_3}(\theta \mid m_3) f_{M_3}(m_3) \end{aligned}$$

where $\mu_{M_2}^{'} = (\mu_{M_2}^1, \ldots, \mu_{M_2}^K)$ and $\lambda_{M_2} = diag(\lambda_{M_2}^1, \ldots, \lambda_{M_2}^K)$. ■

1.10 Appendix B: Deferred Acceptance Algorithm and the Definition of Sharp Samples

The student-proposing Deferred Acceptance (DA) algorithm assigns the exam school offers as follows:

- Round 1: Applicants are considered for a seat in their most preferred exam school. Each exam schools rejects the lowest-ranking applicants in excess of its capacity. The rest of the applicants are provisionally admitted.
- Round $k > 1$: Applicants rejected in Round $k-1$ are considered for a seat in their next most preferred exam school. Each exam schools considers these applicants together with the provisionally admitted applicants from Round $k-1$ and rejects the lowest-ranking students in excess of its capacity. The rest of the students are provisionally admitted.

The algorithm terminates once either all applicants are assigned an offer from one of the exam schools or all applicants with no offer are rejected by every exam school in their preference ordering. This produces an admissions cutoff for each exam school that is given by the lowest rank among applicants admitted to the school. By definition none of the applicants with a ranking below this cutoff are admitted to this school. On the other hand, applicants with a rank at or above this cutoff are admitted to either this school or a more preferred exam school depending on their position relative to the admissions cutoffs for these schools.

The DA algorithm-based admissions process implies that only a subset of the applicants to a given exam school that clear the admissions cutoff are admitted to this school. There are three ways in which an applicant can be admitted to exam school s given the admissions cutoffs:

1. Exam school s is the applicant's 1st choice, and she clears the admissions cutoff.
2. The applicant does not clear the admissions cutoff for her 1st choice, exam school s is her 2nd choice, and she clears the admissions cutoff.
3. The applicant does not clear the admissions cutoff for her 1st or 2nd choice, exam school s is her 3rd choice, and she clears the admissions cutoff.

However, it possible to define for each exam school a sharp sample that consist of applicants who are admitted to this school if and only if they clear the admissions cutoff (Abdulkadiroglu, Angrist, and Pathak, 2014). The sharp sample for exam school s is the union of the following three subsets of applicants:

1. Exam school s is the applicant's 1st choice.
2. The applicant does not clear the admissions cutoff for her 1st choice, and exam school s is her 2nd choice.
3. The applicant does not clear the admissions cutoff for her 1st or 2nd choice, and exam school s is her 3rd choice.

Note that each applicant is included in the sharp sample for at least one exam school (the exam school they listed as their first choice), but an applicant can be included in the sharp sample for more than one exam school. For instance, an applicant who does not clear the admissions cutoff for any of the exam schools is included in the sharp samples for all three schools.

1.11 Appendix C: Identification of the Parametric Latent Factor Model

In this appendix I discuss moment conditions that give identification of the parametric measurement and latent outcome models. I ignore the presence of covariates in this discussion as this can be handled straightforwardly by conditioning on them throughout.

Identification of the Measurement Model

Under the parametric measurement model specified in Section 1.4 the mean and covariances of the measures can be written as

$$\begin{aligned}
E\left[M_1^k\right] &= \mu_{\theta_k} \\
E\left[M_2^k\right] &= \mu_{M_2^k} + \lambda_{M_2^k}\mu_{\theta_k} \\
E\left[M_3^k\right] &= \mu_{M_3^k} + \lambda_{M_3^k}\mu_{\theta_k} \\
Cov\left[M_1^k, M_2^k\right] &= \lambda_{M_2^k}\sigma^2_{\theta_k} \\
Cov\left[M_1^k, M_3^k\right] &= \lambda_{M_3^k}\sigma^2_{\theta_k} \\
Cov\left[M_2^k, M_3^k\right] &= \lambda_{M_2^k}\lambda_{M_3^k}\sigma^2_{\theta_k} \\
Cov\left[M_1^E, M_1^M\right] &= \sigma_{\theta_E\theta_M}
\end{aligned}$$

for $k = E, M$. From these equations one can solve for μ_{θ_k}, $\sigma^2_{\theta_k}$, $\sigma_{\theta_E\theta_M}$, $\mu_{M_2^k}$ and $\lambda_{M_2^k}$ that are given by

$$\begin{aligned}
\mu_{\theta_k} &= E\left[M_1^k\right] \\
\lambda_{M_2^k} &= \frac{Cov\left[M_2^k, M_3^k\right]}{Cov\left[M_1^k, M_3^k\right]} \\
\lambda_{M_3^k} &= \frac{Cov\left[M_3^k, M_2^k\right]}{Cov\left[M_1^k, M_2^k\right]} \\
\mu_{M_2^k} &= E\left[M_2^k\right] - \lambda_{M_2^k}\mu_{\theta_k} \\
\mu_{M_3^k} &= E\left[M_3^k\right] - \lambda_{M_3^k}\mu_{\theta_k} \\
\sigma^2_{\theta_k} &= \frac{Cov\left[M_1^k, M_2^k\right]}{\lambda_{M_2^k}} \\
\sigma_{\theta_E\theta_M} &= Cov\left[M_1^E, M_1^M\right]
\end{aligned}$$

for $k = E, M$, provided that $Cov\left[M_1^k, M_2^k\right], Cov\left[M_1^k, M_3^k\right] \neq 0$, $k = E, M$.

Furthermore, the conditional means and covariances of the measures can be written as

$$\begin{aligned} E\left[M_1^k \mid M_2^k\right] &= E\left[\theta_k \mid M_2^k\right] \\ E\left[M_1^k \mid M_3^k\right] &= E\left[\theta_k \mid M_3^k\right] \\ Cov\left[M_1^k, M_3^k \mid\mid M_2^k\right] &= \lambda_{M_3^k} Var\left[\theta_k \mid M_2^k\right] \\ Cov\left[M_1^k, M_2^k \mid\mid M_3^k\right] &= \lambda_{M_2^k} Var\left[\theta_k \mid M_3^k\right] \end{aligned}$$

for $k = E, M$. From these equations one can solve for $E\left[\theta_k \mid M_2^k\right]$, $E\left[\theta_k \mid M_3^k\right]$, $Var\left[\theta_k \mid M_2^k\right]$, and $Var\left[\theta_k \mid M_3^k\right]$ that are given by

$$\begin{aligned} E\left[\theta_k \mid M_2^k\right] &= E\left[M_1^k \mid M_2^k\right] \\ E\left[\theta_k \mid M_3^k\right] &= E\left[M_1^k \mid M_3^k\right] \\ Var\left[\theta_k \mid M_2^k\right] &= \frac{Cov\left[M_1^k, M_3^k \mid\mid M_2^k\right]}{\lambda_{M_3^k}} \\ Var\left[\theta_k \mid M_3^k\right] &= \frac{Cov\left[M_1^k, M_2^k \mid\mid M_3^k\right]}{\lambda_{M_2^k}} \end{aligned}$$

for $k = E, M$.

Finally, the conditional variances of the measures can be written as

$$\begin{aligned} Var\left[M_1^k \mid M_2^k\right] &= E\left[Var\left[M_1^k \mid \theta_k\right] \mid M_2^k\right] + Var\left[E\left[M_1^k \mid \theta_k\right] \mid M_2^k\right] \\ &= E\left[\exp\left(2\left(\gamma_{M_1^k} + \delta_{M_1^k}\right)\theta_k\right) \mid M_2^k\right] + Var\left[\theta_k \mid M_2^k\right] \\ &= \exp\left(2\left(\gamma_{M_1^k} + E\left[\theta_k \mid M_2^k\right]\delta_{M_1^k} + Var\left[\theta_k \mid M_2^k\right]\delta_{M_1^k}^2\right)\right) + Var\left[\theta_k \mid M_2^k\right] \\ Var\left[M_2^k \mid M_3^k\right] &= E\left[Var\left[M_2^k \mid \theta_k\right] \mid M_3^k\right] + Var\left[E\left[M_2^k \mid \theta_k\right] \mid M_3^k\right] \\ &= E\left[\exp\left(2\left(\gamma_{M_2^k} + \delta_{M_2^k}\right)\theta_k\right) \mid M_3^k\right] + \lambda_{M_2^k}^2 Var\left[\theta_k \mid M_3^k\right] \\ &= \exp\left(2\left(\gamma_{M_2^k} + E\left[\theta_k \mid M_3^k\right]\delta_{M_2^k} + Var\left[\theta_k \mid M_3^k\right]\delta_{M_2^k}^2\right)\right) + \lambda_{M_2^k}^2 Var\left[\theta_k \mid M_{23}^k\right] \\ Var\left[M_3^k \mid M_2^k\right] &= E\left[Var\left[M_3^k \mid \theta_k\right] \mid M_2^k\right] + Var\left[E\left[M_3^k \mid \theta_k\right] \mid M_2^k\right] \\ &= E\left[\exp\left(2\left(\gamma_{M_3^k} + \delta_{M_3^k}\right)\theta_k\right) \mid M_2^k\right] + \lambda_{M_3^k}^2 Var\left[\theta_k \mid M_2^k\right] \\ &= \exp\left(2\left(\gamma_{M_3^k} + E\left[\theta_k \mid M_2^k\right]\delta_{M_3^k} + Var\left[\theta_k \mid M_2^k\right]\delta_{M_3^k}^2\right)\right) + \lambda_{M_3^k}^2 Var\left[\theta_k \mid M_{23}^k\right] \end{aligned}$$

which can be further modified as

$$\begin{aligned} \frac{1}{2}\log\left(Var\left[M_1^k \mid M_2^k\right] - Var\left[\theta_k \mid M_2^k\right]\right) &= \gamma_{M_1^k} + E\left[\theta_k \mid M_2^k\right]\delta_{M_1^k} + Var\left[\theta_k \mid M_2^k\right]\delta_{M_1^k}^2 \\ \frac{1}{2}\log\left(Var\left[M_2^k \mid M_3^k\right] - \lambda_{M_2^k}^2 Var\left[\theta_k \mid M_3^k\right]\right) &= \gamma_{M_2^k} + E\left[\theta_k \mid M_3^k\right]\delta_{M_2^k} + Var\left[\theta_k \mid M_3^k\right]\delta_{M_2^k}^2 \\ \frac{1}{2}\log\left(Var\left[M_3^k \mid M_2^k\right] - \lambda_{M_3^k}^2 Var\left[\theta_k \mid M_2^k\right]\right) &= \gamma_{M_3^k} + E\left[\theta_k \mid M_2^k\right]\delta_{M_3^k} + Var\left[\theta_k \mid M_2^k\right]\delta_{M_3^k}^2 \end{aligned}$$

for $k = E, M$.

Thus, the parameters $\gamma_{M_1^k}$, $\delta_{M_1^k}$, $\gamma_{M_2^k}$, $\delta_{M_2^k}$, $\gamma_{M_3^k}$, and $\delta_{M_3^k}$ can be solved from

$$
\begin{aligned}
&\begin{bmatrix} \frac{1}{2}\log\left(Var\left[M_1^k \mid M_2^k = m_1\right] - Var\left[\theta_k \mid M_2^k = m_1\right]\right) \\ \frac{1}{2}\log\left(Var\left[M_1^k \mid M_2^k = m_2\right] - Var\left[\theta_k \mid M_2^k = m_2\right]\right) \end{bmatrix} \\
&= \begin{bmatrix} \gamma_{M_1^k} + E\left[\theta_k \mid M_2^k = m_1\right]\delta_{M_1^k} + Var\left[\theta_k \mid M_2^k = m_1\right]\delta_{M_1^k}^2 \\ \gamma_{M_1^k} + E\left[\theta_k \mid M_2^k = m_2\right]\delta_{M_1^k} + Var\left[\theta_k \mid M_2^k = m_2\right]\delta_{M_1^k}^2 \end{bmatrix} \\
&\begin{bmatrix} \frac{1}{2}\log\left(Var\left[M_2^k \mid M_3^k = m_1\right] - \lambda_{M_2^k}^2 Var\left[\theta_k \mid M_3^k = m_1\right]\right) \\ \frac{1}{2}\log\left(Var\left[M_2^k \mid M_3^k = m_2\right] - \lambda_{M_2^k}^2 Var\left[\theta_k \mid M_3^k = m_2\right]\right) \end{bmatrix} \\
&= \begin{bmatrix} \gamma_{M_2^k} + E\left[\theta_k \mid M_3^k = m_1\right]\delta_{M_2^k} + Var\left[\theta_k \mid M_3^k = m_1\right]\delta_{M_2^k}^2 \\ \gamma_{M_2^k} + E\left[\theta_k \mid M_3^k = m_2\right]\delta_{M_2^k} + Var\left[\theta_k \mid M_3^k = m_2\right]\delta_{M_2^k}^2 \end{bmatrix} \\
&\begin{bmatrix} \frac{1}{2}\log\left(Var\left[M_3^k \mid M_2^k = m_1\right] - \lambda_{M_3^k}^2 Var\left[\theta_k \mid M_2^k = m_1\right]\right) \\ \frac{1}{2}\log\left(Var\left[M_3^k \mid M_2^k = m_2\right] - \lambda_{M_3^k}^2 Var\left[\theta_k \mid M_2^k = m_2\right]\right) \end{bmatrix} \\
&= \begin{bmatrix} \gamma_{M_3^k} + E\left[\theta_k \mid M_2^k = m_1\right]\delta_{M_3^k} + Var\left[\theta_k \mid M_2^k = m_1\right]\delta_{M_3^k}^2 \\ \gamma_{M_3^k} + E\left[\theta_k \mid M_2^k = m_2\right]\delta_{M_3^k} + Var\left[\theta_k \mid M_2^k = m_2\right]\delta_{M_3^k}^2 \end{bmatrix},
\end{aligned}
$$

provided that the matrices

$$
\begin{bmatrix} 1 & E\left[\theta_k \mid M_2^k = m_1\right] + 2Var\left[\theta_k \mid M_2^k = m_1\right]\delta_{M_1^k} \\ 1 & E\left[\theta_k \mid M_2^k = m_2\right] + 2Var\left[\theta_k \mid M_2^k = m_2\right]\delta_{M_1^k} \end{bmatrix}
$$

$$
\begin{bmatrix} 1 & E\left[\theta_k \mid M_3^k = m_1\right] + 2Var\left[\theta_k \mid M_3^k = m_1\right]\delta_{M_2^k} \\ 1 & E\left[\theta_k \mid M_3^k = m_2\right] + 2Var\left[\theta_k \mid M_3^k = m_2\right]\delta_{M_2^k} \end{bmatrix}
$$

$$
\begin{bmatrix} 1 & E\left[\theta_k \mid M_2^k = m_1\right] + 2Var\left[\theta_k \mid M_2^k = m_1\right]\delta_{M_3^k} \\ 1 & E\left[\theta_k \mid M_2^k = m_2\right] + 2Var\left[\theta_k \mid M_2^k = m_2\right]\delta_{M_3^k} \end{bmatrix}
$$

are of full rank.

Identification of the Latent Outcome Models

Under the parametric latent outcome model specified in Section 1.4 the conditional expectation of an outcome Y can be written as

$$
\begin{aligned}
E\left[Y \mid M, Z\right] &= E\left[Y\left(S\left(Z\right)\right) \mid M, Z\right] \\
&= \alpha_{Y(S(Z))} + \beta_{Y(S(Z))}^{E} E\left[\theta_E \mid M, Z\right] + \beta_{Y(S(Z))}^{M} E\left[\theta_M \mid M, Z\right]
\end{aligned}
$$

for $Z = 0, 1, 2, 3$. Thus, the parameters $\alpha_{Y(S(Z))}$, $\beta^E_{Y(S(Z))}$, and $\beta^M_{Y(S(Z))}$ can be solved from

$$\begin{bmatrix} E\left[h\left(Y,S\right) \mid M = m_1^Z, Z\right] \\ E\left[h\left(Y,S\right) \mid M = m_2^Z, Z\right] \\ E\left[h\left(Y,S\right) \mid M = m_3^Z, Z\right] \end{bmatrix} = \begin{bmatrix} 1 & E\left[\theta_E \mid M = m_1^Z, Z\right] & E\left[\theta_M \mid M = m_1^Z, Z\right] \\ 1 & E\left[\theta_E \mid M = m_2^Z, Z\right] & E\left[\theta_M \mid M = m_2^Z, Z\right] \\ 1 & E\left[\theta_E \mid M = m_3^Z, Z\right] & E\left[\theta_M \mid M = m_3^Z, Z\right] \end{bmatrix} \begin{bmatrix} \alpha_{Y(S(z))} \\ \beta^E_{Y(S(z))} \\ \beta^M_{Y(S(z))} \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} \alpha_{Y(S(Z))} \\ \beta^E_{Y(S(Z))} \\ \beta^M_{Y(S(Z))} \end{bmatrix} = \begin{bmatrix} 1 & E\left[\theta_E \mid M = m_1^Z, Z\right] & E\left[\theta_M \mid M = m_1^Z, Z\right] \\ 1 & E\left[\theta_E \mid M = m_2^Z, Z\right] & E\left[\theta_M \mid M = m_2^Z, Z\right] \\ 1 & E\left[\theta_E \mid M = m_3^Z, Z\right] & E\left[\theta_M \mid M = m_3^Z, Z\right] \end{bmatrix}^{-1} \times \begin{bmatrix} E\left[h\left(Y,S\right) \mid M = m_1^Z, Z\right] \\ E\left[h\left(Y,S\right) \mid M = m_2^Z, Z\right] \\ E\left[h\left(Y,S\right) \mid M = m_3^Z, Z\right] \end{bmatrix}$$

where $m_1^Z, m_2^Z, m_3^Z \in \mathcal{M}^Z$, provided that the matrix

$$\begin{bmatrix} 1 & E\left[\theta_E \mid M = m_1^Z, Z\right] & E\left[\theta_M \mid M = m_1^Z, Z\right] \\ 1 & E\left[\theta_E \mid M = m_2^Z, Z\right] & E\left[\theta_M \mid M = m_2^Z, Z\right] \\ 1 & E\left[\theta_E \mid M = m_3^Z, Z\right] & E\left[\theta_M \mid M = m_3^Z, Z\right] \end{bmatrix}$$

is of full rank. The identification of $\alpha_{D_s(Z)}$, $\beta^E_{D_s(Z)}$, and $\beta^M_{D_s(Z)}$, $Z = 0, 1, 2, 3$, is analogous.

Chapter 2

Wanna Get Away? RD Identification of Exam School Effects Away from the Cutoff

(Joint work with Joshua Angrist)

> Both the tie-breaking experiment and the regression-discontinuity analysis are particularly subject to the external validity limitation of selection-X interaction in that the effect has been demonstrated only for a very narrow band of talent, i.e., only for those at the cutting score... Broader generalizations involve the extrapolation of the below-X fit across the entire range of X values, and at each greater degree of extrapolation, the number of plausible rival hypotheses becomes greater.
>
> – Donald T. Campbell and Julian Stanley (*1963; Experimental and Quasi-Experimental Designs for Research*)

2.1 Introduction

In a regression discontinuity (RD) framework, treatment status changes discontinuously as a function of an underlying covariate, often called the running variable. Provided conditional mean functions for potential outcomes given the running variable are reasonably smooth, changes in outcome distributions at the assignment cutoff must be driven by discontinuities in the likelihood of treatment. RD identification comes from a kind of virtual random assignment, where small and presumably serendipitous variation in the running variable manipulates treatment. On the other hand, because the running variable is usually related to

outcomes, claims for unconditional "as-if random assignment" are most credible for samples near the point of discontinuity. RD methods need not identify causal effects for larger and perhaps more representative groups of subjects. Our epigraph suggests this point was no less apparent to RD's inventors than to today's nonparametricians.

A recent study of causal effects at Boston's selective public schools – known as "exam schools" – highlights the possibly local and potentially limiting nature of RD findings. Boston exam schools choose their students based on an index that combines admissions test scores with a student's grade point average (GPA). Abdulkadiroglu, Angrist, and Pathak (2014) use parametric and non-parametric RD estimators to capture the causal effects of exam school attendance for applicants with index values in the neighborhood of admissions cutoffs. In this case, nonparametric RD compares students just to the left and just to the right of each cutoff. For most of these marginal students, the resulting estimates suggest that exam school attendance does little to boost achievement.[1] But applicants who only barely manage to gain admission to, say, the highly selective Boston Latin School, might be unlikely to benefit from an advanced exam school curriculum. Stronger applicants who qualify more easily may get more from an elite public school education. Debates over affirmative action also focus attention on inframarginal applicants, including some who stand to gain seats and some who stand to lose their seats should affirmative action considerations be brought in to the admissions process.[2]

Motivated by the question of how exam school attendance affects achievement for inframarginal applicants, this paper tackles the theoretical problem of RD identification for applicants other than those in the immediate neighborhood of admissions cutoffs. Our first tack extrapolates parametric models for conditional mean functions estimated to the left and right of cutoffs. As noted by Angrist and Pischke (2009), in a parametric framework, extrapolation is easy.

As it turns out, functional-form-based estimation procedures fail to produce compelling results for the empirical question that motivates our theoretical inquiry. The resulting estimates of exam school effects away from the cutoff are mostly imprecise and sensitive to the polynomial used for extrapolation, with or without the implicit weighting induced by a nonparametric bandwidth. We therefore turn to a conditional independence argument that exploits a key feature of most RD assignment mechanisms: treatments is assigned as a deterministic function of a single observed covariate, the running variable. The association between running variable and outcome variables is therefore the *only* source of omitted variables bias in RD estimates. If, for example, the running variable were randomly assigned, or otherwise made independent of potential outcomes, we could ignore it and analyze data from RD designs as if from a randomized trial.

The special nature of RD assignment leads us to a conditional independence assumption (CIA) that identifies causal effects by conditioning on covariates besides the running variable, with an eye to eliminating the relationship between running variable and outcomes. It's not always possible to find such good controls,

[1] In an RD study of New York exam schools, Dobbie and Fryer (2013) similarly find little evidence of gains for admitted applicants at the cutoff.

[2] Duflo, Dupas, and Kremer (2011) use a combination of RD and a randomized trial to document treatment effect heterogeneity as a function of the running variable in a study of tracking in Kenyan elementary schools.

of course, but, as we show below, a straightforward statistical test isolates promising candidates. As an empirical matter, we show that conditioning on baseline scores and demographic variables largely eliminates the relationship between running variables and test score outcomes for 9th grade applicants to Boston exam schools, though not for 7th grade applicants (for whom the available controls are not as good). These results lay the foundation for a matching strategy that identifies causal effects for inframarginal 9th applicants.

Our estimates of effects away from the cutoff are mostly in line with RD estimates of causal effects at the cutoff. In particular, away-from-the-cutoff estimates suggest BLS attendance has little effect on either math or English achievement, while the O'Bryant school may generate some gains, especially in English Language Arts (ELA). The ELA gains for successful O'Bryant applicants approach one-fifth of a standard deviation. Perhaps surprisingly, therefore, those who seem most likely to gain from any expansion in exam school seats are relatively weak applicants who currently fail to gain admission to Boston's least selective exam school. Ultra-high ability applicants, that is, BLS applicants who easily clear the threshold for Boston's most selective public school, are likely to do well with our without the benefit of a BLS experience, at least as far as standardized test scores go.

2.2 Causal Effects at Boston Exam Schools

Boston's three exam schools serve grades 7-12. The high-profile Boston Latin School (BLS), which enrolls about 2,400 students, is the oldest American high school, founded in 1635. BLS is a model for other exam schools, including New York's well-known selective high schools. The second oldest Boston exam school is Boston Latin Academy (BLA), formerly Girls' Latin School. Opened in 1877, BLA first admitted boys in 1972 and currently enrolls about 1,700 students. The John D. O'Bryant High School of Mathematics and Science (formerly Boston Technical High) is Boston's third exam school; O'Bryant opened in 1893 and now enrolls about 1,200 students.

The Boston Public School (BPS) system spans a wide range of peer achievement. Like many urban students elsewhere in the U.S., Boston exam school applicants who fail to enroll in an exam school end up at schools with average SAT scores well below the state average, in this case, at schools close to the 5th percentile of the distribution of school averages in the state. By contrast, O'Bryant's average SAT scores fall near the 40th percentile of the state distribution of averages, a big step up from the overall BPS average, but not elite in an absolute sense. Successful Boston BLA applicants find themselves at a school with average scores around the 80th percentile of the distribution of school means, while the average SAT score at BLS is the fourth highest among public schools in Massachusetts.

Between 1974 and 1998, Boston exam schools reserved seats for minority applicants. Though quotas are no longer in place, the role of race in exam school admissions continues to be debated in Boston and is the subject of ongoing litigation in New York. Our CIA-driven matching strategy is used here to answer two questions about the most- and least-selective of Boston's three exam schools; both questions are motivated

by the contemporary debate over affirmative action in exam school admissions. Specifically, we ask:

1. How would inframarginal low-scoring applicants to O'Bryant, Boston's least selective exam school, do if they were lucky enough to find seats at O'Bryant in spite of falling a decile or more below today's O'Bryant cutoff? In other words, what if poorly qualified O'Bryant applicants now at a regular BPS school were given the opportunity to attend O'Bryant?

2. How would inframarginal high-scoring applicants to BLS, Boston's most selective exam school and one of the most selective in the country, fare if their BLS offers were withdrawn in spite of the fact that they qualify easily by today's standards? In other words, what if highly qualified applicants now at BLS had to settle for BLA?

The first of these questions addresses the impact of exam school attendance on applicants who currently fail to make the cut for any school but might do so with minority preferences restored or exam school seats added in an effort to boost minority enrollment. The second question applies to applicants like Julia McLaughlin, whose 1996 lawsuit ended racial quotas at Boston exam schools. McLaughlin was offered a seat at BLA, but sued for a seat at BLS, arguing, ultimately successfully, that she was kept out of BLS solely by unconstitutional racial quotas. The thought experiment implicit in our second question sends high-scoring BLS students like McLaughlin back to BLA.

2.2.1 Data

The data used here merge BPS enrollment and demographic information with Massachusetts Comprehensive Assessment System (MCAS) scores. MCAS tests are taken each spring, typically in grades 3-8 and 10. Baseline (i.e., pre-application) scores for grade 7 applicants are from 4th grade. Baseline English scores for 9th grade applicants come from 8th grade math and 7th grade ELA tests (the 8th grade English exam was introduced in 2006). We lose some applicants with missing baseline scores. Scores were standardized by subject, grade, and year to have mean zero and unit variance in the BPS population.

Data on student enrollment, demographics and test scores were combined with the BPS exam school applicant file. This file records applicants' current grade and school enrolled, applicants' preference ordering over exam schools, and applicants' Independent Schools Entrance Exam (ISEE) test scores, along with each exam schools' ranking of its applicants as determined by ISEE scores and GPA. These school-specific rankings become the exam school running variables in our setup.

Our initial analysis sample includes BPS-enrolled students who applied for exam school seats in 7th grade from 1999-2008 or in 9th grade from 2001-2007. We focus on applicants enrolled in BPS at the time of application (omitting private school students) because we're interested in how an exam school education compares to regular district schools. Moreover, private school applicants are much more likely to remain outside the BPS district and hence out of our sample if they fail to get an exam school offer. Applicants who apply to transfer from one exam school to another are also omitted.

2.2.2 Exam School Admissions

The sharp CIA-based estimation strategy developed here is predicated on the notion that exam school offers are a deterministic function of exam school running variables. Exam school running variables are constructed by ranking a weighted average of ISEE scores and applicants' GPAs at the time of application. In practice, however, Boston exam school offers take account of student preferences over schools as well as their ISEE scores and GPAs. Students list up to three exam schools for which they wish to be considered, in order of preference. Admissions offers are determined by a student-proposing deferred acceptance (DA) algorithm, using student preferences and school-specific running variables as inputs. The DA matching process complicates our RD analysis because it loosens the direct link between running variables and admissions offers. As in Abdulkadiroglu, Angrist, and Pathak (2014), our econometric strategy begins by constructing analysis samples that restore a deterministic link between exam school offers and running variables, so that offers are sharp around admissions cutoffs. A description of the manner in which these *sharp samples* are constructed appears in the appendix.[3]

The sharp RD treatment variable is an offer dummy, denoted D_{ik}, indicating applicants offered a seat at school k, determined separately as a function of rank for applicants in each school-specific sharp sample. For the purposes of empirical work, school-specific ranks are centered and scaled to produce the following running variable:

$$r_{ik} = \frac{100}{N_k} \times (\tau_k - c_{ik}), \tag{2.1}$$

where N_k is the total number of students who ranked school k (not the number in the sharp sample). Scaled school-specific ranks, r_{ik}, equal zero at the cutoff rank for school k, with positive values indicating students who ranked and qualified for admission at that school. Absent centering, scaled ranks give applicants' percentile position in the distribution of applicants to school k. Within sharp samples, we focus on a window limited to applicants with running variables no more than 20 units (percentiles) away from the cutoff. For qualified 9th grade applicants at BLS, this is non-binding since the BLS cutoff is closer to the top of the 9th grade applicant distribution than the .8 quantile.

In sharp samples, offers are determined by the running variable, but not all offers are accepted. This can be seen in Figures 2-1a and 2-1b, which plot school-specific offer and enrollment rates around O'Bryant and BLS admissions cutoffs. Specifically, the figures show conditional means for sharp sample applicants in a one-unit binwidth, along with a conditional mean function smoothed using local linear regression (LLR).[4]

[3]Instead of defining sharp samples, a dummy for threshold crossing (qualification) can be used to instrument fuzzy offers. The extension of our CIA approach to fuzzy designs is discussed in Section 2.5, below. The construction of sharp sample produces an asymptotic efficiency gain, however, since those in the sharp sample are compliers in a setup that uses qualification as an instrument for offers (this is implied by results in Frolich (2007) and Hong and Nekipelov (2010), which show that the ability to predict compliance reduces the semiparametric efficiency bound for local average treatment effects.)

[4]For school k, data in the estimation window were used to construct estimates of $\hat{\mathrm{E}}[y_i|r_{ik}]$, where y_i is the dependent variable and r_{ik} is the running variable. The LLR smoother uses the edge kernel,

$$K_h(r_{ik}) = 1\Big\{\Big|\frac{r_{ik}}{h}\Big| \leq 1\Big\} \cdot \Big(1 - \Big|\frac{r_{ik}}{h}\Big|\Big),$$

where h is the bandwidth. The bandwidth used here is a version of the DesJardins and McCall (2008) bandwidth (hereafter,

As can be seen in Table 2.1, which reports estimates that go with these figures, 72% of 7th graders offered a seat at O'Bryant enroll there, while among 9th grade applicants offered an O'Bryant seat, 66% enroll. Enrollment rates are much higher for those offered a seat at BLS, while many applicants not offered a seat at BLS end up at Boston's second most selective exam school, BLA. At the same time, movement up the ladder of exam school selectivity is associated with dramatic changes in peer composition. This can be seen in Figure 2-2a and 2-2b, which plot peer achievement of applicants' classmates (as measured by baseline MCAS scores), for applicants within 20 percentile points of the O'Bryant and BLS cutoffs.

2.2.3 Results at the Cutoff

As a benchmark, we begin with estimates for marginal applicants. Figures 2-3a and 2-3b show little evidence of gains in 10th grade math scores for 7th grade applicants offered exam school seats. On the other hand, among both 7th and 9th grade applicants, 10th grade ELA scores seem to jump at the O'Bryant cutoff. The figure also hints at an O'Bryant-induced gain in math scores, though only for 9th grade applicants.

Our estimators of the effect of an exam school offer are derived from models for potential outcomes. Let Y_{1i} and Y_{0i} denote potential outcomes in treated and untreated states, with the observed outcome determined by

$$y_i = Y_{0i} + [Y_{1i} - Y_{0i}]D_i.$$

In a parametric setup, the conditional mean functions for potential outcomes given the running variable are modeled as:

$$\begin{aligned} E[Y_{0i}|r_i] &= f_0(r_i) \\ E[Y_{1i}|r_i] &= \rho + f_1(r_i), \end{aligned}$$

using polynomials, $f_j(r_i); j = 0, 1$.

Substituting polynomials in $E[y_i|r_i] = E[Y_{0i}|r_i] + E[Y_{1i} - Y_{0i}|r_i]D_i$, and allowing for the fact that the estimation sample pools data from different test years and application years, the parametric estimating equation for applicant i observed in year t is:

$$y_{it} = \alpha_t + \sum_j \beta_j p_{ij} + \sum_\ell \delta_\ell d_{i\ell} + (1 - D_i)f_0(r_i) + D_i f_1(r_i) + \rho D_i + \eta_{it} \tag{2.2}$$

This model controls for test year effects, denoted α_t, and for application year, indexed by ℓ and indicated by dummies, $d_{i\ell}$. The model also includes a full set of application preference dummies, denoted p_{ij}.[5] The effects of the running variable are controlled by a pair of pth-order polynomials that differ on either side of

DM) studied by Imbens and Kalyanaraman (2012), who derive optimal bandwidths for sharp RD using a mean square-error loss function with a regularization adjustment. The DM smoother (which generates somewhat more stable estimates in our application than the bandwidth Imbens and Kalyanaraman (2012) prefer) is also used to construct nonparametric RD estimates, below.

[5] As explained in the appendix, this controls for applicant-preference-group composition effects in the sharp sample.

the cutoff, specifically:

$$f_j(r_i) = \pi_{1j} r_i + \pi_{2j} r_i^2 + ... + \pi_{pj} r_i^p; \quad j = 0, 1. \tag{2.3}$$

The benchmark estimates set $p = 3$.

Non-parametric RD estimators differ from parametric in three ways. First, they narrow the estimation window when the optimal data-driven bandwidth falls below 20. Non-parametric estimators also use a tent-shaped edge kernel centered at admissions cutoffs, instead of the uniform kernel implicit in parametric estimation. Finally, non-parametric models control for linear functions of the running variable only, omitting higher-order terms. The nonparametric estimating equation is:

$$\begin{aligned} y_{it} &= \alpha_t + \sum_j \beta_j p_{ij} + \sum_\ell \delta_\ell d_{i\ell} + \gamma_0 (1 - D) r_i + \gamma_1 D_i r_i + \rho D_i + \eta_{it} \\ &= \alpha_t + \sum_j \beta_j p_{ij} + \sum_\ell \delta_\ell d_{i\ell} + \gamma_0 r_i + \gamma^* D_i r_i + \rho D_i + \eta_{it} \end{aligned} \tag{2.4}$$

Non-parametric RD estimates come from a kernel-weighted LLR fit of equation (2.4), estimated separately in the sharp sample of applicants to O'Bryant and BLS.

Consistent with the figures, estimates of (2.2) and (2.4), reported in Table 2.2, show little in the way of score gains at BLS. But the non-parametric estimates suggest an O'Bryant offer may boost 10th grade ELA scores for both 7th and 9th grade applicants. Other estimates are either smaller or less precise, though among 9th grade O'Bryant applicants, we see a marginally significant effect on math. Other estimates, not reported here, present a broad picture of small effects on 7th grade exam school applicants tested in 7th and 8th grade (see Abdulkadiroglu, Angrist, and Pathak (2014) for nonparametric estimates of effects on middle school scores.) Results for the 10th grade ELA scores of O'Bryant applicants offer the strongest evidence of an exam school gain.

2.2.4 To Infinity and Beyond: Parametric Extrapolation

The running variable is the star covariate in any RD scene, but the role played by the running variable is distinct from that played by covariates in matching and regression-control strategies. In the latter, we look to comparisons of treated and non-treated observations *conditional* on covariates to eliminate omitted variables bias. As Figure 2-4 highlights, however, in an RD design, there is *no* value of the running variable at which both treatment and control subjects are observed. Nonparametric identification comes from infinitesimal changes in covariate values across the RD cutoff. As a practical matter, however, nonparametric inference procedures compare applicants with covariate values in a small - though not infinitesimal - neighborhood to the left of the cutoff with applicants whose covariate values put them in a small neighborhood to the right. This empirical comparison requires some extrapolation, however modest. Identification of causal effects away from the cutoff requires a more substantial extrapolative leap.

In a parametric setup such as described by (2.2) and (2.3), extrapolation is easy though not necessarily

credible. For any distance, c, we have

$$\rho(c) \equiv E[Y_{1i} - Y_{0i}|r_i = c] = \rho + \pi_1^* c + \pi_2^* c^2 + ... + \pi_p^* c^p, \tag{2.5}$$

where $\pi_1^* = \pi_{11} - \pi_{10}$, and so on. The notation in (2.5) masks the extrapolation challenge inherent in identification away from the cutoff: potential outcomes in the treated state are observed for $r_i = c > 0$, but the value of $E[Y_{0i}|r_i = c]$ for positive c is never seen. The dotted lines in Figure 2-4 show two equally plausible possibilities, implying different causal effects at $r_i = c$. It seems natural to use observations to the left of the cutoff in an effort to pin down functional form, and then extrapolate this to impute $E[Y_{0i}|r_i = c]$. With enough data, and sufficiently well-behaved conditional mean functions, $f_0(c)$ is identified for all values of c, including those never seen in the data. It's easy to see, however, why this approach may not generate robust or convincing findings.

The unsatisfying nature of parametric extrapolation emerges in Figures 2-5a and 2-5b. These figures show observed and imputed counterfactual 10th grade math scores for 7th and 9th grade applicants. Specifically, the figures plot nonparameteric estimates of the observed conditional mean function $E[Y_{0i}|r_i = c]$ for O'Bryant applicants to the left of the cutoff, along with imputed $E[Y_{1i}|r_i = c]$ to the left. Similarly, for BLS applicants, the figures plot nonparametric estimates of observed $E[Y_{1i}|r_i = c]$ for applicants to the right of the cutoff, along with imputed $E[Y_{0i}|r_i = c]$ to the right. The imputations use linear, quadratic, and cubic specifications for $f_j(r_i)$. These models generate a wide range of estimates, especially as distance from the cutoff grows. For instance, the estimated effect of BLS attendance to the right of the cutoff for 9th grade applicants changes sign when the polynomial goes from second to third degree. This variability seems unsurprising and consistent with Campbell and Stanley (1963)'s observation that, "at each greater degree of extrapolation, the number of plausible rival hypotheses becomes greater." On the other hand, given that $f_0(r_i)$ looks reasonably linear for $r_i < 0$ and $f_1(r_i)$ looks reasonably linear for $r_i > 0$, we might have hoped for results consistent with those from linear models, even when the specification allows something more elaborate.

Table 2.3, which reports the estimates and standard errors from the models used to construct the fitted values plotted in Figure 2-5, shows that part of the problem uncovered in the figure is imprecision. Estimates constructed with $p = 3$ are too noisy to be useful at $c = +/- 5$ or higher. Models setting $p = 2$ generate more precise estimates than when $p = 3$, though still fairly imprecise for $c \geq 10$. On the other hand, for very modest extrapolation ($c = 1$), a reasonably consistent picture emerges. Like RD estimates at the cutoff, this slight extrapolation generates small positive estimates at O'Bryant and small negative effects at BLS for both 7th and 9th grade applicants, though few of these estimates are significantly different from zero.[6]

[6]Paralleling Figure 2-5, the estimates in Table 2.3 are from models omitting controls for test year, application year and application preferences. Estimates from models with these controls differ little from those reported in the table.

2.2.4.1 Using Derivatives Instead

Dong and Lewbel (2013) propose an alternative to parametric extrapolation based on the insight that the derivatives of conditional mean functions are nonparametrically identified at the cutoff (a similar idea appears in Section 3.3.2 of DiNardo and Lee, 2011). First-order derivative-based extrapolation exploits the fact that

$$f_j(c) \approx f_j(0) + f_j'(0) \cdot c. \tag{2.6}$$

This approximation can be implemented using a nonparametric estimate of $f_j'(0)$.

The components of (2.6) are estimated consistently by fitting linear models to $f_j(r_i)$ in a neighborhood of the cutoff, using a data-driven bandwidth and slope terms that vary across the cutoff. Specifically, the effect of an offer at cutoff value c can be approximated as

$$\rho(c) \approx \rho + \gamma^* \cdot c, \tag{2.7}$$

with parameters estimated using equation (2.4). The innovation in this procedure relative to LLR estimation of (2.4) is in the interpretation of the interaction term, γ^*. Instead of a bias-reducing nuisance parameter, γ^* is seen in this context as identifying a derivative that facilitates extrapolation. As a practical matter, the picture that emerges from derivative-based extrapolation of exam school effects is similar to that shown in Figure 2-5.

2.3 Call in the CIA

RD designs take the mystery out of treatment assignment. In sharp samples of applicants to Boston exam schools, we know that exam school offers are determined by

$$D_i = 1[r_i > 0].$$

This signal feature of the RD design implies that failure to control for r_i is the only source of omitted variables bias in estimates of the causal effect of D_i.

Armed with precise knowledge of the source of omitted variables bias, we propose to identify causal effects by means of a conditional independence argument. In sharp samples, Boston exam school offers are determined by measures of past achievement, specifically ISEE scores and students' GPAs. But these are not the only lagged achievement measures available. In addition to demographic variables that are highly predictive of achievement, we observe pre-application scores on MCAS tests taken in 4th grade and, for high school applicants, in 7th or 8th grade. Conditioning on this rich and relevant set of controls may serve to break the link between running variables and outcomes.[7]

[7]Cook (2008) credits Goldberger (1972a) and Goldberger (1972b) for the observation that when treatment status is de-

To formalize this identification strategy, we gather the set of available controls in a covariate vector, x_i. Our conditional independence assumption (CIA) asserts that:

CONDITIONAL INDEPENDENCE ASSUMPTION (CIA)

$$E[Y_{ji}|r_i, x_i] = E[Y_{ji}|x_i];\ j = 0, 1$$

In other words, potential outcomes are assumed to be mean-independent of the running variable conditional on x_i. We also require treatment status to vary conditional on x_i:

COMMON SUPPORT

$$0 < P[D_i = 1|x_i] < 1\ a.s.$$

The CIA and common support assumptions identify any counterfactual average of interest. For example, the average of Y_{0i} to the right of the cutoff is:

$$E[Y_{0i}|D_i = 1] = E\{E[Y_{0i}|x_i, D_i = 1]|D_i = 1\} = E\{E[y_i|x_i, D_i = 0]|D_i = 1\}, \tag{2.8}$$

while the average treatment effect on the treated is identified by a matching-style estimand:

$$E[Y_{1i} - Y_{0i}|D_i = 1] = E\{E[y_i|x_i, D_i = 1] - E[y_i|x_i, D_i = 0]|D_i = 1\}.$$

2.3.1 Testing and Bounding

Just as with conventional matching strategies (as in, for example, Heckman, Ichimura, and Todd (1998) and Dehejia and Wahba (1999)), the CIA assumption invoked here breaks the link between treatment status and potential outcomes, opening the door to identification of a wide range of average causal effects. In this case, however, the prior information inherent in an RD design is also available to guide our choice of the conditioning vector, x_i. Specifically, by virtue of the conditional independence relation implied by the CIA, we have:

$$E[Y_{1i}|r_i, x_i, r_i > 0] = E[Y_{1i}|x_i] = E[Y_{1i}|x_i, r_i > 0],$$

so we should expect that

$$E[y_i|r_i, x_i, D_i = 1] = E[y_i|x_i, D_i = 1], \tag{2.9}$$

to the right of the cutoff. Likewise, the CIA also implies:

$$E[Y_{0i}|r_i, x_i, r_i < 0] = E[Y_{0i}|x_i] = E[Y_{0i}|x_i, r_i < 0],$$

termined solely by a pre-treatment test score, regression control for pre-treatment scores eliminates omitted variables bias. Goldberger credits Barnow (1972) and Lord and Novick (1972) for similar insights.

suggesting we look for

$$E[y_i|r_i, x_i, D_i = 0] = E[y_i|x_i, D_i = 0], \tag{2.10}$$

to the left of the cutoff.

Regressions of outcomes on x_i and the running variable on either side of the cutoff provide a simple test for (2.9) and (2.10). Mean independence is stronger than regression independence, of course, but regression testing procedures can embed flexible models that approximate nonlinear conditional mean functions. In practice, simple regression-based tests seem likely to provide the most useful specification check since such tests are likely to reject in the face of any sort of dependence between outcomes and running variable, while more elaborate specifications with many free parameters may lack the power to detect violations.[8]

Concerns about power notwithstanding, the CIA is demanding and may be hard to satisfy. A weaker and perhaps more realistic version limits the range of running variable values for which the CIA is maintained. This weaker bounded conditional independence assumption asserts that the CIA holds only over a limited range:

BOUNDED CONDITIONAL INDEPENDENCE ASSUMPTION (BCIA)

$$E[Y_{ji}|r_i, x_i, \mid r_i \mid < d] = E[Y_{ji}|x_i, \mid r_i \mid < d];\ j = 0, 1$$

Bounded CIA says that potential outcomes are mean-independent of the running variable conditional on x_i, but only in a d-neighborhood of the cutoff. Testing BCIA, we look for

$$E[y_i|r_i, x_i, 0 < r_i < d] = E[y_i|x_i, 0 < r_i < d] \tag{2.11}$$

to the right of the cutoff, and

$$E[y_i|r_i, x_i, -d < r_i < 0] = E[y_i|x_i, -d < r_i < 0] \tag{2.12}$$

to the left of the cutoff.

At first blush, the BCIA evokes nonparametric RD identification in that it leads to estimation of casual effects inside an implicit bandwidth around the cutoff. An important distinction, however, is the absence of any promise to make the d-neighborhood smaller as the sample size grows. Likewise, BCIA requires no choice of bandwidth or local polynomial smoothers with an eye to bias-variance trade-offs. Rather, the *largest* value of d that appears to satisfy BCIA defines the playing field for CIA-based estimation.

Beyond providing an opportunistic weakening of the CIA, the BCIA assumption allows us to avoid bias from counterfactual composition effects as distance from the cutoff grows. Moving, say, to the left of the BLS cutoff, BLS applicants start to fall below the BLA cutoff as well, thereby changing the relevant

[8]Fan and Li (1996), Lavergne and Vuong (2000), Ait-Sahalia, Bickel, and Stoker (2001), and Angrist and Kuersteiner (2011) develop nonparametric conditional independence tests.

counterfactual from BLA to O'Bryant for BLS applicants not offered a seat there. The resulting change in Y_{0i} (where potential outcomes are indexed against BLS offers) is likely to be correlated with the BLS running variable with or without conditioning on x_i. To argue otherwise requires the distinction between BLA and O'Bryant to be of no consequence. BCIA avoids the resulting composition bias by requiring that we not extrapolate too far to the left of the BLS cutoff when looking at BLS applicants.

2.3.2 Alternative Assumptions and Approaches

A weaker alternative to the CIA asserts conditional independence between average causal effects and the running variable, instead of between potential outcomes and the running variable. This leads to a Conditional Effect Ignorability (CEI) assumption, similar to that introduced by Angrist and Fernandez-Val (2010) in an instrumental variables setting. For our purposes, CEI can be described as follows:

CONDITIONAL EFFECT IGNORABILITY (CEI)

$$E[Y_{1i} - Y_{0i}|r_i, x_i] = E[Y_{1i} - Y_{0i}|x_i]$$

In an RD context, CEI means that, conditional on x_i, we can ignore the running variable when computing average causal effects, even if potential outcomes are not individually mean-independent of the running variable.[9]

CEI has much of the identifying power of the CIA. For example, given CEI, the effect of treatment on the treated can be written as:

$$E[Y_{1i} - Y_{0i}|D_i = 1] = E\{E[y_i|x_i, r_i = 0^+] - E[y_i|x_i, r_i = 0^-]|D_i = 1\}, \tag{2.13}$$

where $E[y_i|x_i, r_i = 0^+]$ and $E[y_i|x_i, r_i = 0^-]$ denote right- and left-hand limits of conditional-on-x_i expectation functions for outcomes at the cutoff. In other words, this CEI assumption identifies causal effects away from the cutoff by averaging a set of nonparametrically identified conditional-on-covariates effects at the cutoff.

In practice, CIA-based estimates seem likely to be more useful than those derived from equation (2.13). For one thing, not being limited to identification near the cutoff, CIA-based estimation uses much more data. Second, CEI relies on the ability to find a fair number of observations near the cutoff for all relevant covariate values, a tall order in many applications. Finally, CEI is harder to assess. CEI implies that the derivative of the conditional average treatment effect given covariates should be zero at the cutoff; as noted by Dong and Lewbel (2013), this derivative is non-parametrically identified (and given by the interaction term in the nonparametric estimating equation, (2.4)). In practice, however, samples large enough for reliable nonparametric estimates of conditional mean functions can be expected to generate inconclusive estimates

[9]Lewbel (2007) invokes a similar assumption in a setup using exclusion restrictions to correct for classification error in treatment status.

of derivatives. Not surprisingly, therefore, our experiments with CEI estimators for Boston exam school applicants failed to produce estimates that seem precise enough to be useful.

Battistin and Rettore (2008) also consider matching estimates in an RD setting, though they don't exploit an RD-specific conditional independence condition. Rather, in the spirit of Lalonde (1986), Battistin and Rettore validate a generic matching estimator by comparing non-parametric RD estimates with conventional matching estimates constructed at the cutoff. They argued that when matching and RD produce similar results at the cutoff, matching seems worth exploring away from the cutoff as well.

Other related discussions of RD identification away from the cutoff include DiNardo and Lee (2011) and Lee and Lemieux (2010), both of which note that the local interpretation of nonparametric RD estimates can be relaxed by treating the running variable as random rather than conditioning on it. In this view, observed running variable values are the realization of a non-degenerate stochastic process that assigns values to individuals of an underlying type. Each type contributes to local-to-cutoff average treatment effects in proportion to that type's likelihood of being represented at the cutoff. Since "type" is an inherently latent construct, this random running variable interpretation doesn't seem to offer concrete guidance as to how causal effects might change away from the cutoff. In the spirit of this notion of latent conditioning, however, we might model running variables and the conditioning variables in our CIA assumption as noisy measures of a single underlying ability measure. In ongoing work, Rokkanen (2014) explores RD models where identification is based on this sort of latent factor ignorability in a structural econometric framework.

Finally, moving in a different direction, Jackson (2010) outlines an extrapolation approach that identifies inframarginal effects at exam schools in Trinidad and Tobago by exploiting the fact that students with the same running variable (a test score) can end up at different schools, depending on their preferences. Jackson (2010) identifies effects away from the cutoff by differences-in-differences style contrasts between infra-marginal high- and low-scoring applicants with different rankings. Cook and Wing (2013) explore a similar idea, offering supportive Monte Carlo evidence for a hybrid differences-in-differences/RD approach.

2.3.3 CIA-based Estimators

We economize on notation by omitting explicit conditioning on running variable values falling in the $[-d, d]$ interval; expectations in this section should be understood to be conditional on the largest value of d that satisfies BCIA. Where relevant, the constant c is assumed to be no bigger than d in absolute value.

At specific running variable values, the CIA leads to the following matching-style estimand:

$$\begin{aligned} &E[Y_{1i} - Y_{0i} | r_i = c] = \\ &\quad E\{E[y_i | x_i, D_i = 1] - E[y_i | x_i, D_i = 0] | r_i = c\} \end{aligned} \tag{2.14}$$

Alternately, on the right-hand side of the cutoff, we might consider causal effects averaged over all positive

values up to c, a bounded effect of treatment on the treated:

$$E[Y_{1i} - Y_{0i}|0 < r_i \leq c] = \\ E\{E[y_i|x_i, D_i = 1] - E[y_i|x_i, D_i = 0]|0 < r_i \leq c\} \tag{2.15}$$

Paralleling this on the left, the bounded effect of treatment on the non-treated is:

$$E[Y_{1i} - Y_{0i}| - c \leq r_i < 0] = \\ E\{E[y_i|x_i, D_i = 1] - E[y_i|x_i, D_i = 0]| - c \leq r_i < 0\} \tag{2.16}$$

We consider two estimators of (2.14), (2.15) and (2.16). The first is a linear reweighting estimator discussed by Kline (2011). The second is a version of the Hirano, Imbens, and Ridder (2003) propensity score estimator based on Horvitz and Thompson (1952). We also use the estimated propensity score to document common support, as in Dehejia and Wahba's (1999) pioneering propensity score study of the effect of a training program on earnings.

Kline's reweighting estimator begins with linear models for conditional means, which can be written:

$$\begin{aligned} E[y_i|x_i, D_i = 0] &= x_i'\beta_0 \\ E[y_i|x_i, D_i = 1] &= x_i'\beta_1 \end{aligned} \tag{2.17}$$

Linearity is not really restrictive since the parametrization for $x_i'\beta_j$ can be rich and flexible. Substituting in (2.14), we have

$$E[Y_{1i} - Y_{0i}|r_i = c] \\ = (\beta_1 - \beta_0)'E[x_i|r_i = c], \tag{2.18}$$

with similar expressions based on (2.15) and (2.16).

Let $\lambda(x_i) \equiv E[D_i|x_i]$ denote the propensity score. Our propensity score weighting estimator begins with the observation that the CIA implies

$$\begin{aligned} E\left[\frac{y_i(1 - D_i)}{1 - \lambda(x_i)}|x_i\right] &= E[Y_{0i}|x_i] \\ E\left[\frac{y_i D_i}{\lambda(x_i)}|x_i\right] &= E[Y_{1i}|x_i] \end{aligned}$$

Bringing these expressions inside a single expectation and over a common denominator, the treatment effect

on the treated for those with $0 < r_i < c$ is given by

$$E[Y_{1i} - Y_{0i}|0 < r_i \leq c] = E\left\{\frac{y_i[D_i - \lambda(x_i)]}{\lambda(x_i)\left[1 - \lambda(x_i)\right]} \cdot \frac{P[0 < r_i \leq c|x_i]}{P[0 < r_i \leq c]}\right\}. \tag{2.19}$$

Similar formulas give the average effect for non-treated applicants and average effects at specific, possibly narrow, ranges of running variable values. The empirical counterpart of (2.19) requires a model for the probability $P[0 < r_i \leq c|x_i]$ as well as for $\lambda(x_i)$. It seems natural to use the same parameterization for both. Note also that if $c = d$, the estimand in (2.19) simplifies to

$$E[Y_{1i} - Y_{0i}|D_i = 1] = E\left\{\frac{y_i[D_i - \lambda(x_i)]}{[1 - \lambda(x_i)]E[D_i]}\right\},$$

as in Hirano, Imbens, and Ridder (2003).[10]

2.4 The CIA in Action at Boston Exam Schools

We start by testing BCIA in estimation windows that set d equal to 10, 15, and 20. Regressions used for testing control for baseline test scores along with indicators of special education status, limited English proficiency, eligibility for free or reduced price lunch, race (black/Asian/Hispanic) and sex, as well as indicators for test year, application year and application preferences. Baseline score controls for 7th grade applicants include 4th grade math and ELA scores, while for 9th grade applicants, baseline scores include 7th grade ELA scores and 8th grade math scores.

CIA test results, reported in Table 2.4, show that conditioning fails to eliminate the relationship between running variables and potential outcomes for 7th grade applicants; most of the estimated coefficients are significantly different from zero for both 10th grade math and ELA scores. At the same time, test results for 9th grade applicants seem promising. Most test statistics (that is, running variable coefficient estimates) for 9th grade applicants are smaller than the corresponding statistics for 7th grade applicants, and only one is significantly different from zero (this is for math scores to the left of the BLS cutoff in the $d = 20$ window). It should be noted, however, that few 9th grade applicants fall to the right of the BLS cutoff. CIA tests for BLS applicants with $D_i = 1$ are forgiving because the sample for this group is small.[11]

We complement formal CIA testing with a graphical tool motivated by an observation in Lee and Lemieux (2010): in a randomized trial using a uniformly distributed random number to determine treatment assignment, this number becomes the running variable for an RD design. The relationship between outcomes and running variable should be flat, however, except possibly for a jump at the quantile cutoff which de-

[10]The expectations and conditioning here refer to distributions in the sharp sample of applicants for each school. Thus, treatment effects on the treated are for treated applicants in a school-k sharp sample. When the estimand targets average effects at specific $r_i = c$, as opposed to over an interval, the probabilities $P[r_i = c|x_i]$ and $P[r_i = c]$ needed for (2.19) become densities. Note also that the estimand in (2.19) can be written $E[\omega_{1i}y_i - \omega_{0i}y_i]$, where $E[\omega_{0i}] = E[\omega_{1i}] = 1$. As noted by Imbens (2004), however, this need not hold in finite samples. We therefore normalize the sum of these weights to be 1.

[11]The unchanging sample size to the right of the BLS cutoff as d shrinks reflects the high BLS admissions threshold for 9th grade applicants: the $d = 10$ limit isn't binding for BLS on the right.

termines proportion treated. Our CIA assumption implies this same pattern. Figure 2-6 therefore plots 10th grade math and ELA residuals constructed by partialing out x_i against running variables in a $d = 20$ window. The figure shows conditional means for all applicants in one-unit binwidths, along with conditional mean functions smoothed using local linear regression. Consistent with the test results reported in Table 2.4, Figure 2-6 shows a strong positive relationship between outcome residuals and running variables for 7th grade applicants. For 9th grade applicants, however, the relationship between outcome residuals and running variables is essentially flat, except perhaps for ELA scores in the BLS sample.

The difference in CIA test results for 7th and 9th grade applicants may be due to the fact that baseline scores for 9th grade applicants come from a grade closer to the outcome test grade than for 7th grade applicants. In combination with demographic control variables and 4th grade scores, 7th or 8th grade MCAS scores do a good job of eliminating the running variable from 9th graders' conditional mean functions for 10th grade scores. By contrast, the most recent baseline test scores available for 7th grade applicants are from 4th grade tests.[12] In view of the results in Table 2.4 and Figure 2-6, the CIA-based estimates that follow are for 9th grade applicants only.

Columns 1-4 of Table 2.5 report linear reweighting estimates of average treatment effects. These are estimates of $E[Y_{1i} - Y_{0i}|0 < r_i < d]$ for BLS applicants and $E[Y_{1i} - Y_{0i}| - d < r_i < 0]$ for O'Bryant applicants, in samples that set d equal to 10, 15, and 20. The estimand for BLS is

$$\begin{aligned} &E[Y_{1i} - Y_{0i}|0 < r_i \leq d] \\ &\quad = (\beta_1 - \beta_0)' E[x_i|0 < r_i \leq d], \end{aligned} \tag{2.20}$$

while that for O'Bryant is

$$\begin{aligned} &E[Y_{1i} - Y_{0i}| - d \leq r_i < 0] \\ &\quad = (\beta_1 - \beta_0)' E[x_i| - d \leq r_i < 0], \end{aligned} \tag{2.21}$$

where β_0 and β_1 are defined in (2.17). The BLS estimand is an average effect of treatment on the treated, since treated observations in the estimation window must have positive running variables. Likewise, the O'Bryant estimand is an average effect of treatment on the non-treated.

As with RD estimates at the cutoff, the CIA results in Table 2.5 show no evidence of a BLS achievement boost. At the same time, results for inframarginal unqualified O'Bryant applicants offer some evidence of gains, especially in ELA. The math estimates range from $.09\sigma$ when $d = 10$ to $.16\sigma$ when $d = 20$, though the estimate effect for $d = 10$ is only marginally significantly different from zero. Linear reweighting results for the ELA scores of O'Bryant applicants are clear cut, however, ranging from $.18\sigma$ to $.2\sigma$ and significantly different from zero for each choice of d. The CIA estimates are remarkably consistent with the corresponding

[12]The addition of quadratic and cross-subject interaction terms in baseline scores fails to improve CIA test results for 7th grade applicants.

RD estimates at the cutoff: compare, for example, the CIA estimates in columns 1 and 3 of Table 2.5 to the nonparametric O'Bryant RD estimates at the cutoff of $.13\sigma$ (SE=.07) in math and $.18\sigma$ (SE=.07) for ELA, shown in column 3 of Table 2.2.

Figure 2-7 completes the picture on effects away from the cutoff by plotting linear reweighting estimates of $E[Y_{1i}|r_i = c]$ and $E[Y_{0i}|r_i = c]$ for all values of c in the $[-20, 20]$ interval. To the left of the O'Bryant cutoff, the estimates of $E[Y_{0i}|r_i = c]$ are fitted values from regression models for observed outcomes, while the estimates of $E[Y_{1i}|r_i = c]$ are implicitly an extrapolation and labelled accordingly. To the right of the BLS cutoff, the estimates of $E[Y_{1i}|r_i = c]$ are fitted values while the estimates of $E[Y_{0i}|r_i = c]$ are an extrapolation. The conditional means in this figure were constructed by plugging individual values of x_i into (2.17) and smoothing the results using local linear regression.[13] The figure presents a picture consistent with that arising from the estimates in Table 2.5. In particular, the extrapolated BLS effects are small (for ELA) or noisy (for math), while the O'Bryant extrapolation reveals a remarkably stable gain in ELA scores away from the cutoff. The extrapolated effect of O'Bryant offers on math scores appears to increase modestly as a function of distance from the cutoff, a finding probed further below.

2.4.1 Propensity Score Estimates

CIA-based estimation of the effect of exam school offers seems like a good setting for propensity score methods, since the conditioning set includes multiple continuously distributed control variables. These features of the data complicate full covariate matching. Our logit model for the propensity score uses the same control variables and parametrization as were used to construct the tests in Table 2.4 and the linear reweighting estimates in columns 1-4 of Table 2.5.[14]

The estimated propensity score distributions for admitted and rejected applicants exhibit a substantial degree of overlap. This is documented in Figure 2-8, which plots the histogram of estimated scores for treated and control observations above and below a common horizontal axis. Not surprisingly, the larger sample of O'Bryant applicants generates more overlap than the sample for highly selective BLS. Most score values for untreated O'Bryant applicants fall below about .6. Each decile in the O'Bryant score distribution contains at least a few treated observations; above the first decile, there appear to be more than enough for accurate inference. By contrast, few untreated BLS applicants have covariate values for which a BLS offer is highly likely. We should therefore expect the BLS counterfactual to be estimated less precisely than that for O'Bryant.

It's also worth noting that because the sample contains no BLS controls with propensity score values above .8 (or .9 in one window), the BLS estimates fail to reflect outcomes for applicants with admissions probabilities above this value. Figure 2-8 documents other noteworthy features of conditional-on-score comparisons: the O'Bryant treatment effect on the non-treated implicitly compares the many non-treated applicants with low

[13] Smoothing here uses the edge kernel with Stata's default bandwidth.

[14] Propensity score models for the smaller sample of BLS applicants omit test date and application preference dummies.

scores to the fewer (though still plentiful) treated O'Bryant applicants with scores in this range; the BLS treatment effect on the treated compares a modest number of treated applicants, more or less uniformly distributed across score values, with corresponding untreated observations, of which many more are low-scoring than high.

The propensity-score-weighted estimates reported in columns 5-8 of Table 2.5 are remarkably consistent with the linear reweighting estimates shown in columns 1-4 of the table. In particular, the estimates here suggest most BLS students would do no worse if they had had to go to BLA instead, while low scoring O'Bryant applicants might enjoy substantial gains in ELA were they offered a seat at O'Bryant. At the same time, the propensity score estimates for BLS applicants reported in columns 6 and 8 are highly imprecise. These BLS estimates are not only much less precise than the corresponding O'Bryant estimates, the standard errors here are two-four times larger than those generated by linear reweighting for the same samples. Linear reweighting looks like an attractive procedure in this context.[15]

2.5 Fuzzy CIA Models

Estimates of the effect of O'Bryant offers on the ELA scores of 9th grade applicants are reasonably stable as distance from the cutoff grows. By contrast, the estimated effect of O'Bryant offers on math scores appears to increase as window width or distance from the cutoff increases. In a window of width 10, for example, estimated O'Bryant math effects are only marginally significantly different from zero, while the estimate in a window of width 20 is almost twice as large and significant (at $.16\sigma$ with a standard error of $.05\sigma$). Taken at face value, this finding suggests that the weakest 9th grade applicants stand to gain the most from O'Bryant admission, an interesting substantive finding. Omitted variables bias (failure of CIA) seems unlikely to explain this pattern since the relevant conditional independence tests, reported in columns 1 and 5 of Table 2.4, show no violations of CIA.

An alternative explanation for the pattern of O'Bryant math estimates plotted in Figure 2-7 begins with the observation that exam school offers affect achievement by facilitating exam school enrollment. Assuming, as seems plausible, that exam school offers affect outcomes solely through enrollment (that is, other causal channels, such as peer effects, are downstream to enrollment), the estimates in Table 2.5 can be interpreted as the reduced form for an instrumental variables (IV) procedure in which exam school enrollment is the endogenous variable. The magnitude of reduced form comparisons is easier to interpret when the relevant first stage estimates scale these effects. If the first stage changes as a function of the running variable, comparisons of reduced form estimates across running variable values are meaningful only after rescaling. In principle, IV methods make the appropriate adjustment. A question that arises here, however, is how to interpret IV estimates constructed under the CIA in a world of heterogeneous potential outcomes, where the

[15]The standard errors reported in this table use a bootstrap with 500 replications. Bootstrap standard errors provide asymptotically valid confidence intervals for estimators like (2.19) since, as note by Hirano, Imbens, and Ridder (2003), the propensity-score-weighting estimator is asymptotically linear. As noted at the end of Section 2.3.1, estimates based on CEI instead of the CIA are imprecise. Still, the general pattern is similar, suggesting positive effects at O'Bryant only.

average causal effects identified by IV potentially vary with the running variable.

We estimate and interpret the causal effects of exam school enrollment by adapting the dummy treatment/dummy instrument framework outlined in Abadie (2003). This framework allows for unrestricted treatment effect heterogeneity in potentially nonlinear IV models with covariates. The starting point is notation for potential treatment assignments, W_{0i} and W_{1i}, indexed against the instrument, in this case, exam school offers indicated by D_i. Thus, W_{0i} indicates (eventual) exam school enrollment among those not offered a seat, while W_{1i} indicates (eventual) exam school enrollment among those offered a seat. Observed enrollment status is

$$W_i = W_{0i}(1 - D_i) + W_{1i}D_i.$$

The core identifying assumption in our IV setup is a generalized version of CIA:

GENERALIZED CONDITIONAL INDEPENDENCE ASSUMPTION (GCIA)

$$(Y_{0i}, Y_{1i}, W_{0i}, W_{1i}) \perp\!\!\!\perp r_{ik} \mid x_i$$

GCIA can be assumed to hold in a $d-$neighborhood of the cutoff as with BCIA. We also maintain the common support assumption given in Section 2.3.

The GCIA generalizes simple CIA in three ways. First, GCIA imposes full independence instead of mean independence; this seems innocuous since any behavioral or assignment mechanism satisfying the latter is likely to satisfy the former. Second, along with potential outcomes, the pair of potential treatment assignments (W_{0i} and W_{1i}) is taken to be conditionally independent of the running variable. Finally, GCIA requires joint independence of all outcome and assignment variables, while the CIA in Section 2.3 requires only marginal (mean) independence. Again, its hard to see why we'd have the latter without the former.

2.5.1 Fuzzy Identification

As in Section 2.3.3, the expectations in this section should be understood to be conditional on the largest value of d that satisfies GCIA.

2.5.1.1 Local Average Treatment Effects

In a local average treatment effects (LATE) framework with Bernoulli treatment and Bernoulli instruments, the subset of compliers consists of individuals whose treatment status can be changed by changing the instrument. This group is defined here by $W_{1i} > W_{0i}$. A key identifying assumption in the LATE framework is monotonicity: the instrument can only shift treatment one way. Assuming that the instrument D_i satisfies monotonicity with $W_{1i} \geq W_{0i}$, and that for some i the inequality is strong, so there is a first-stage, the LATE theorem (Imbens and Angrist, 1994) tells us that

$$\frac{E[y_i|D_i = 1] - E[y_i|D_i = 0]}{E[W_i|D_i = 1] - E[W_i|D_i = 0]} = E[Y_{1i} - Y_{0i}|W_{1i} > W_{0i}]$$

In other words, a simple Wald-type IV estimator captures average causal effects on exam school applicants who enroll when they receive an offer but not otherwise.

Abadie (2003) generalizes the LATE theorem by showing that the expectation of any measurable function of treatment, covariates, and outcomes is identified for compliers. This result facilitates IV estimation using a wide range of causal models, including nonlinear models such as those based on the propensity score. Here, we adapt the Abadie (2003) result to a fuzzy RD setup that identifies causal effects away from the cutoff. This requires a conditional first stage, described below:

CONDITIONAL FIRST STAGE

$$P[W_{1i} = 1|x_i] > P[W_{0i} = 1|x_i]\, a.s$$

Given GCIA, common support, monotonicity, and a conditional first stage, the following identification result can be established (see the appendix for details):

THEOREM 1 (FUZZY CIA EFFECTS)

$$E[Y_{1i} - Y_{0i}|W_{1i} > W_{0i}, 0 < r_i \leq c]$$
$$= \frac{1}{P[W_{1i} > W_{0i}|0 < r_i \leq c]} E\left\{\psi(D_i, x_i)\frac{P[0 < r_i \leq c|x_i]}{P[0 < r_i \leq c]} y_i\right\} \tag{2.22}$$
$$for \quad \psi(D_i, x_i) \equiv \frac{D_i - \lambda(x_i)}{\lambda(x_i)[1 - \lambda(x_i)]} \tag{2.23}$$

Estimators based on (2.22) capture causal effects for compliers with running variable values falling into any range over which there's common support.[16]

At first blush, it's not immediately clear how to estimate the conditional compliance probability, $P[W_{1i} > W_{0i}|0 < r_i \leq c]$, appearing in the denominator of (2.22). Because everyone to the right of the cutoff is offered treatment, there would seem to be no data available to estimate compliance rates conditional on $0 < r_i \leq c$ (in the original LATE framework, the IV first stage measures the probability of compliance). Paralleling an argument in Abadie (2003), however, the appendix shows that

$$P[W_{1i} > W_{0i}|0 < r_i \leq c] = E\left\{\kappa\left(W_i, D_i x_i\right)\frac{P\left[0 < r_i \leq c \mid x_i\right]}{P\left[0 < r_i \leq c\right]}\right\} \tag{2.24}$$

where

$$\kappa\left(W_i, D_i x_i\right) = 1 - \frac{W_i\left(1 - D_i\right)}{1 - \lambda(x_i)} - \frac{\left(1 - W_i\right)D_i}{\lambda(x_i)}.$$

[16]The weighting function in the numerator is much like that used to construct average treatment effects in Hirano, Imbens, and Ridder (2003) and Abadie (2005). Extensions of this theorem along the lines suggested by Theorem 3.1 in Abadie (2003) identify the marginal distributions of Y_{0i} and Y_{1i}.

2.5.1.2 Average Causal Response

The causal framework leading to Theorem 1 is limited to Bernoulli endogenous variables. For some applicants, however, the exam school treatment is mediated by years of attendance rather than a simple go/no-go decision. We develop a fuzzy CIA estimator for ordered treatments by adapting a result from Angrist and Imbens (1995). The ordered treatment framework relies on potential outcomes indexed against an ordered treatment, w_i. In this context, potential outcomes are denoted by Y_{ji} when $w_i = j$, for $j = 0, 1, 2, ..., J$. We assume also that potential treatments, w_{1i} and w_{0i}, satisfy monotonicity with $w_{1i} \geq w_{0i}$ and generate a conditional first stage:

$$E[w_{1i}|x_i] \neq E[w_{0i}|x_i]$$

The Angrist and Imbens (1995) Average Causal Response (ACR) theorem describes the Wald IV estimand as follows:

$$\frac{E[y_i \mid D_i = 1] - E[y_i \mid D_i = 0]}{E[w_i \mid D_i = 1] - E[w_i \mid D_i = 0]} = \sum_j \nu_j E[Y_{ji} - Y_{j-1,i} \mid w_{1i} \geq j > w_{0i}]$$

where

$$\begin{aligned} \nu_j &= \frac{P[w_{1i} \geq j > w_{0i}]}{\sum_\ell P[w_{1i} \geq \ell > w_{0i}]} \\ &= \frac{P[w_i \leq j \mid D_i = 0] - P[w_i \leq j \mid D_i = 1]}{E[w_i \mid D_i = 1] - E[w_i \mid D_i = 0]} \end{aligned}$$

Wald-type IV estimators therefore capture a weighted average of the average causal effect of increasing w_i from $j-1$ to j, for compliers whose treatment intensity is moved by the instrument from below j to above j. The weights are given by the impact of the instrument on the cumulative distribution function (CDF) of the endogenous variable at each intensity.

The GCIA assumption allows us to establish a similar result in a fuzzy RD setup with an ordered treatment. The following is shown in the appendix:

THEOREM 2 (FUZZY AVERAGE CAUSAL RESPONSE)

$$\begin{aligned} &\frac{E\{E[y_i \mid D_i = 1, x_i] - E[y_i \mid D_i = 0, x_i] \mid 0 < r_i \leq c\}}{E\{E[w_i \mid D_i = 1, x_i] - E[w_i \mid D_i = 0, x_i] \mid 0 < r_i \leq c\}} \\ &= \sum_j \nu_{jc} E[Y_{ji} - Y_{j-1,i} \mid w_{1i} \geq j > w_{0i}, 0 < r_i \leq c] \end{aligned} \tag{2.25}$$

where

$$\nu_{jc} = \frac{P[w_{1i} \geq j > w_{0i} \mid 0 < r_i \leq c]}{\sum_\ell P[w_{1i} \geq \ell > w_{0i} \mid 0 < r_i \leq c]} \tag{2.26}$$

This theorem says that a Wald-type estimator constructed by averaging covariate-specific first-stages and re-

duced forms can be interpreted as a weighted average causal response for compliers with running variable values in the desired range. The incremental average causal response, $E\left[Y_{ji} - Y_{j-1,i} \mid w_{1i} \geq j > w_{0i}, 0 < r_i \leq c\right]$, is weighted by the conditional probability the instrument moves the ordered treatment through the point at which the incremental effect is evaluated.

In practice, we estimate the left hand side of (2.25) by fitting linear models with covariate interactions to the reduced form and first stage. The resulting estimation procedure adapts Kline (2011) to an ordered treatment and works as follows: estimate conditional linear reduced forms interacting D_i and x_i; use these estimates to construct the desired average reduced form effect as in (2.20) and (2.21); divide by a similarly constructed average first stage.[17] The same procedure can be used to estimate (2.25) for a Bernoulli treatment like W_i, in which case the average causal response identified by Theorem 2 becomes the average causal effect identified by Theorem 1 (though the corresponding estimates won't be algebraically the same unless the propensity score model used under Theorem 1 is linear).

2.5.2 Fuzzy Estimates

As with the sharp estimates discussed in Section (2.4), fuzzy enrollment effects are estimated for applicants to the left of the O'Bryant cutoff and to the right of the BLS cutoff, in windows setting d equal to 10, 15 and 20. The enrollment first stage changes remarkably little as distance from the cutoff grows. This can be seen in columns 1-4 of Table 2.6, which report estimates of the effect of exam school offers on exam school enrollment, constructed separately for O'Bryant and BLS applicants using equation (2.24). The propensity score model is the same as that used to construct the estimates in Table 2.5 (Table 2.6 shows separate first stage estimates for the math and ELA samples, as these differ slightly). Given this stable first stage, its unsurprising that estimates of $E[Y_{1i} - Y_{0i}|W_{1i} > W_{0i}, 0 < r_i \leq d]$, reported in columns 5-8 of the table, change little as a function of d. The pattern here is consistent with that in Table 2.5, with small and statistically insignificant effects at BLS, and evidence of large effects at O'Bryant. Estimates of O'Bryant effects on ELA scores range from an impressive gain of $.38\sigma$ when $d = 20$, to a still-substantial $.27\sigma$ when the window is half as wide. The estimated O'Bryant effects on math scores are also considerable, varying from $.17\sigma$ to $.23\sigma$.

The gains for inframarginal applicants who enroll at O'Bryant are perhaps too large to be credible and may therefore signal failure of the underlying exclusion restriction, which channels all causal effects of an exam through an enrollment dummy. Many who start in an exam school drop out, so we'd like to adjust these estimates for years of exam school exposure. We therefore treat years of exam school enrollment as the endogenous variable and estimate the ACR parameter on the right-hand side of equation (2.25), using the modified linear reweighting procedure described above. The covariate parameterization used to construct both reduced form and first stage estimates is the same as that used to construct the sharp estimates in

[17] Specifically, let ϕ_0 be the main effect of D_i and let ϕ_1 be the vector of interactions with x_i in a first stage regression of w_i on D_i, x_i, and $D_i x_i$. The denominator of (2.25) is $\phi_0 + \phi_1' \mu_{xc}$, where $\mu_{xc} = E[x_i | 0 \leq r_i \leq c]$.

Table 2.5.

First stage estimates for years of exam school enrollment, reported in columns 1-4 of Table 2.7, indicate that successful BLS applicants spend about 1.8 years in BLS between application and test date, while successful O'Bryant applicants spend about 1.4 years at O'Bryant between application and test date. The associated ACR estimates, reported in columns 5-8 of the table, are in line with those in Table 2.6, but considerably more precise. For example, the effect of a year of BLS exposure on ELA scores is estimated to be no more than about $.05\sigma$, with a standard error of roughly the same magnitude. This compares with estimates of about the same size in column 8 of Table 2.6, but standard errors for the latter are five or more times larger. The precision gain here would seem to come from linearity of the estimator and not the change in endogenous variable, paralleling precision gains seen in the switch from propensity score to linear reweighting when constructing the sharp estimates in Table 2.5.

ELA estimates for O'Bryant show gains of about $.14\sigma$ per year of exam school exposure, a finding that appears to be more stable across window width than the corresponding dummy enrollment estimates in column 7 of Table 2.6. This comparison suggests that some of the variability seen in the Table 2.6 estimates comes from a failure to adjust for small changes in the underlying first stage for years of enrollment across windows (as can be seen in column 3 of Table 2.7. At the same time, the estimated O'Bryant math gains in column 5 of Table 2.7 still fade in a narrower window, a pattern seen for the O'Bryant math estimates in Tables 2.5 and 2.6.

2.6 Summary and Directions for Further Work

RD estimates of the effect of Boston exam school offers generate little evidence of an achievement gain for most applicants on the margin of admission, but these results need not be relevant for applicants with running variable values well away from admissions cutoffs. This observation motivates RD-inspired identification strategies for causal effects away from the cutoff. Parametric extrapolation seems like a natural first step, but a parametric approach generates unsatisfying estimates of the effects of exam school offers, sensitive to functional form and too imprecise to be useful. We therefore turn to identification strategies based on a conditional independence assumption that focuses on the running variable.

A key insight emerging from the RD framework is that the only source of omitted variables bias is the running variable. Our conditional independence assumption therefore makes the running variable ignorable, that is, independent of potential outcomes, by conditioning on other predictors of outcomes. When the running variable is ignorable, treatment is ignorable. The conditional independence assumption underlying ignorability has strong testable implications that are easily checked in this context. Specifically, the CIA implies that in samples limited to either treated or control observations, regressions of outcomes on the running variable and the covariates supporting CIA should show no running variable effects. A modified or bounded version of the CIA asserts that this conditional independence relation holds only in a neighborhood

of the cutoff.

Among 9th grade applicants to the O'Bryant school and the Boston Latin School, bounded conditional independence appears to hold over a reasonably wide interval. Importantly, the conditioning variables supporting this result include 7th or 8th grade and 4th grade MCAS scores, all lagged versions of the 10th grade outcome variable. Lagged middle school scores in particular seems like a key control, probably because these relatively recent baseline tests are a powerful predictor of future scores. Lagged outcomes are better predictors, in fact, than the running variable itself, which is a composite constructed from applicants' GPAs and a distinct exam school admissions test.

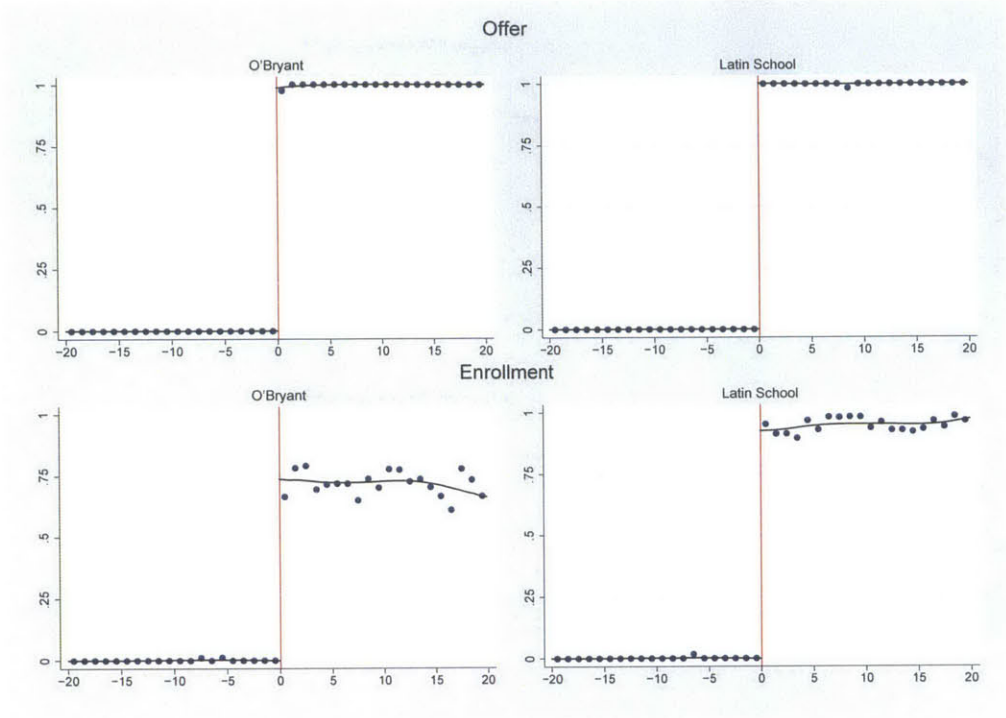
Results based on the CIA suggest that inframarginal high-scoring BLS applicants gain little (in terms of achievement) from BLS attendance, a result consistent with the RD estimates of BLS effects at the cutoff reported in Abdulkadiroglu, Angrist, and Pathak (2014). At the same time, CIA-based estimates using both linear and propensity score models generate robust evidence of strong gains in English for unqualified inframarginal O'Bryant applicants. Evidence of 10th grade grade ELA gains also emerge from the RD estimates of exam school effects reported by Abdulkadiroglu, Angrist, and Pathak (2014), especially for nonwhites. The CIA-based estimates reported here suggest similar gains would likely be observed should the O'Bryant cutoff be reduced to accommodate currently inframarginal high school applicants, perhaps as a result of re-introducing affirmative action considerations in exam school admissions.

We also modify CIA-based identification strategies for fuzzy RD and use this modification to estimate the effects of exam school enrollment and years of exam school attendance, in addition to the reduced form effects of exam school admissions offers. A fuzzy analysis allows us to explore the possibility that changes in reduced form offer effects as a function of the running variable are driven by changes in an underlying first stage for exam school exposure. Interestingly, the fuzzy extension opens the door to identification of causal effects for compliers in RD models for quantile treatment effects. As noted recently by Frandsen, Frolich, and Melly (2012), the weighting approach used by Abadie, Angrist, and Imbens (2002) and Abadie (2003) breaks down in a conventional RD framework because the distribution of treatment status is degenerate conditional on the running variable. By taking the running variable out of the equation, our framework circumvents this problem, a feature we plan to exploit in future work on distributional outcomes.

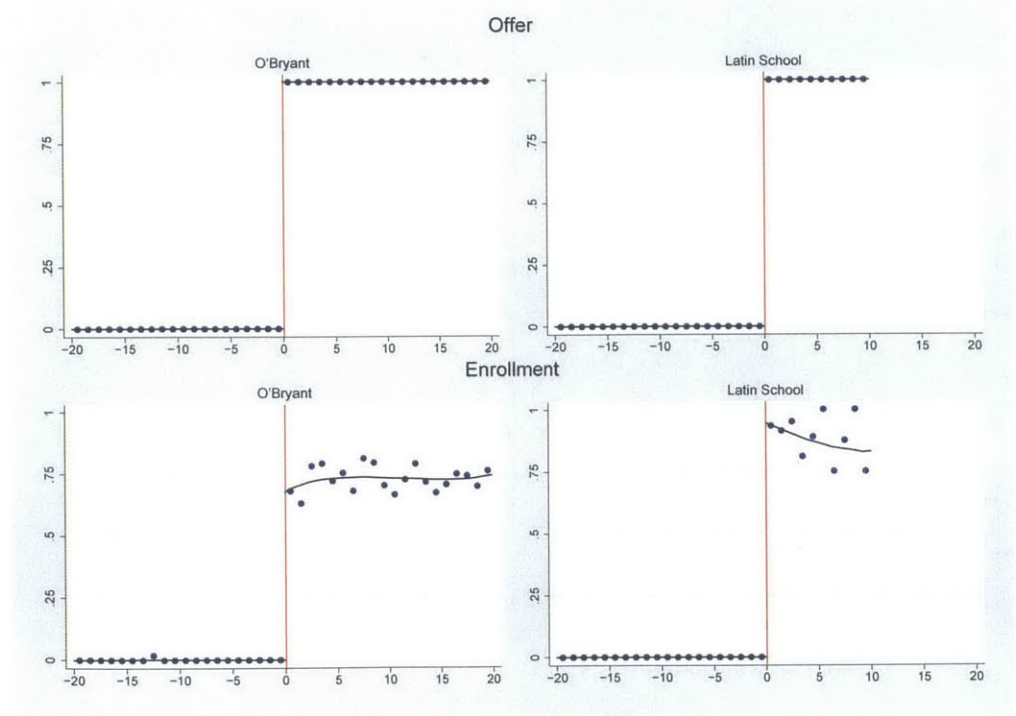
In a parallel and ongoing investigation, Rokkanen (2014) develops identification strategies for RD designs in which the CIA conditioning variable is an unobserved latent factor. Multiple noisy indicators of the underlying latent factor provide the key to away-from-the-cutoff identification in this new context. An important unsolved econometric problem implicit in our empirical strategy is causal inference conditional on a pretest. Estimators that condition on the results of a specification test may have sampling distributions for which conventional asymptotic approximations are poor. Pretesting is a challenging and virtually ubiquitous problem in applied econometrics. It remains to be seen whether recent theoretical progress on the pretesting problem (e.g., Andrews and Guggenberger (2009); Belloni, Chernozhukov, and Hansen (2012)) can be applied fruitfully in this context.

Finally, the mixed results reported here raise the question of what might explain the variation in our estimates across schools. In a pair of recent papers, Abdulkadiroglu, Angrist, Dynarski, Kane, and Pathak (2011) and Angrist, Cohodes, Dynarski, Pathak, and Walters (2013) document large gains at Boston charter high schools when using admissions lotteries to estimate the effects of charter attendance relative to regular district schools. These gains appear to vary inversely with students' baseline achievement, suggesting that the quality of the implicit counterfactual may be an important driver of the treatment effects arising from school choice. The fallback school for most O'Bryant applicants (a regular district school) may have lower value-added than the fallback school for BLS applicants (mostly the BLA exam school), even though the gain in peer quality is larger at the admissions cutoff for the latter. In ongoing work, we're continuing to explore the nexus linking school choice, school quality, and measures of students' baseline ability.

2.7 Figures and Tables

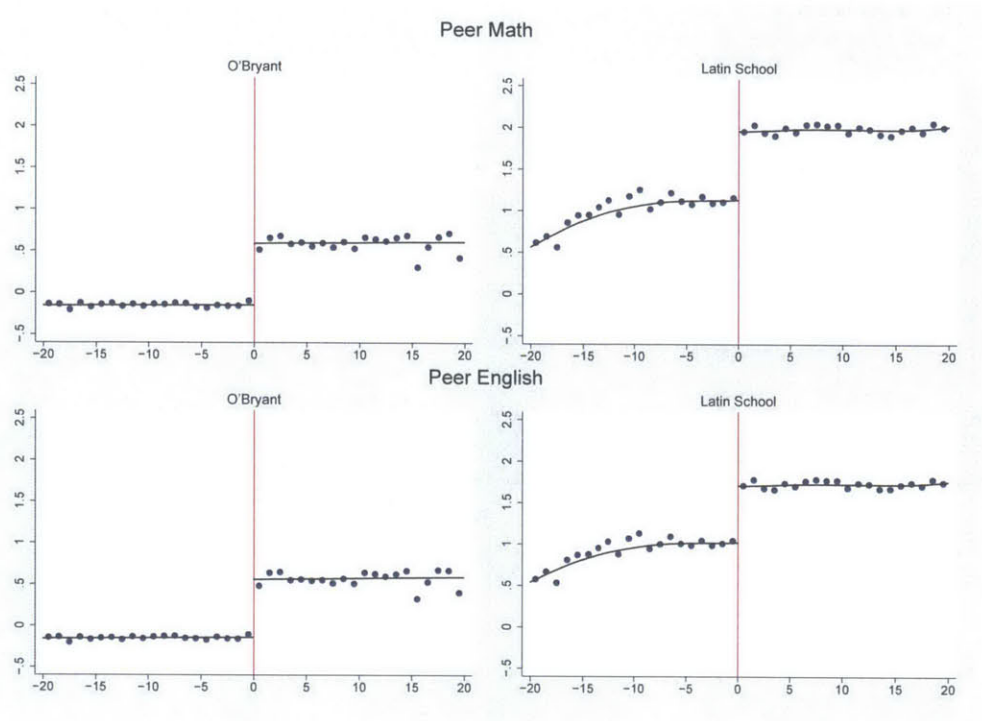


(a) 7th Grade Applicants

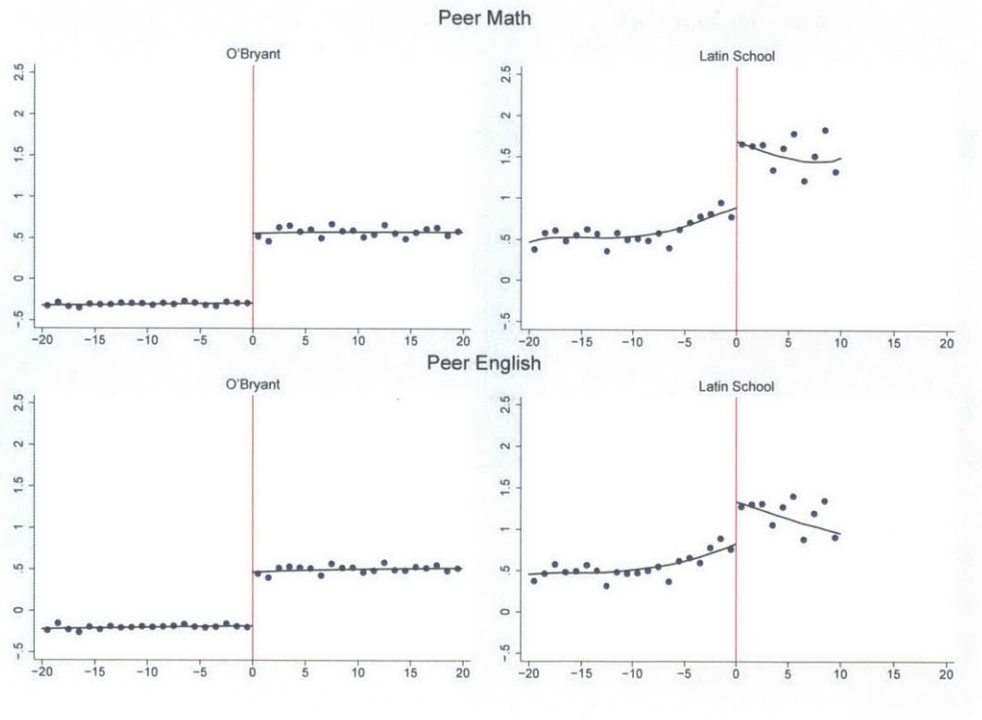


(b) 9th Grade applicants

Figure 2-1: Offer and Enrollment at O'Bryant and Boston Latin School

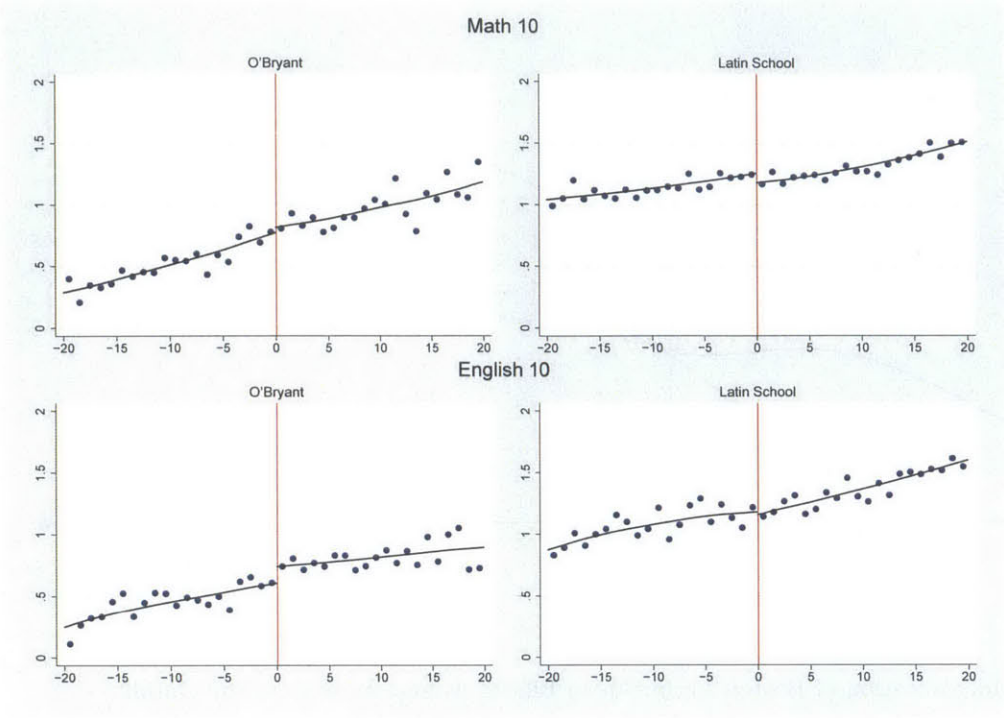


(a) 7th Grade Applicants

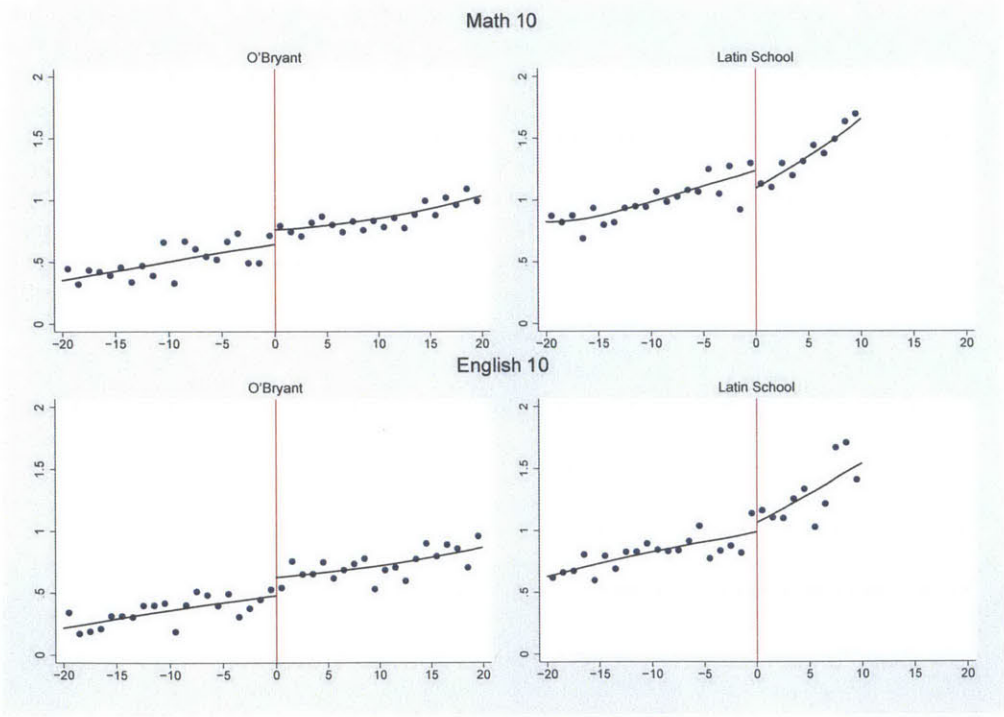


(b) 9th Grade applicants

Figure 2-2: Peer Achievement at O'Bryant and Boston Latin School



(a) 7th Grade Applicants



(b) 9th Grade applicants

Figure 2-3: 10th Grade Math and ELA Scores at O'Bryant and Boston Latin Schools

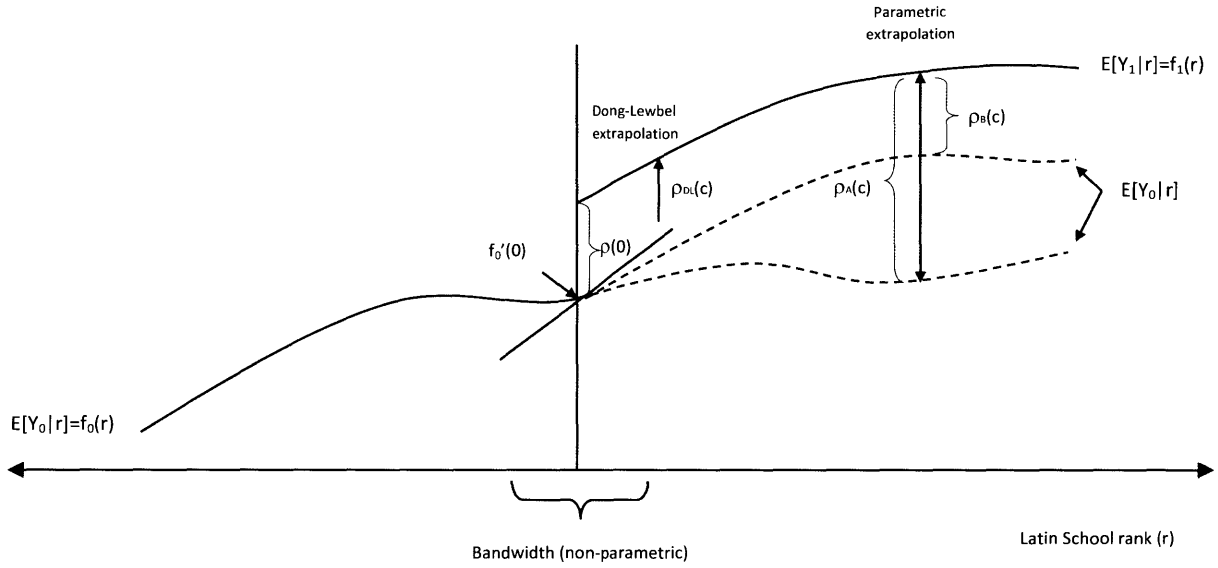
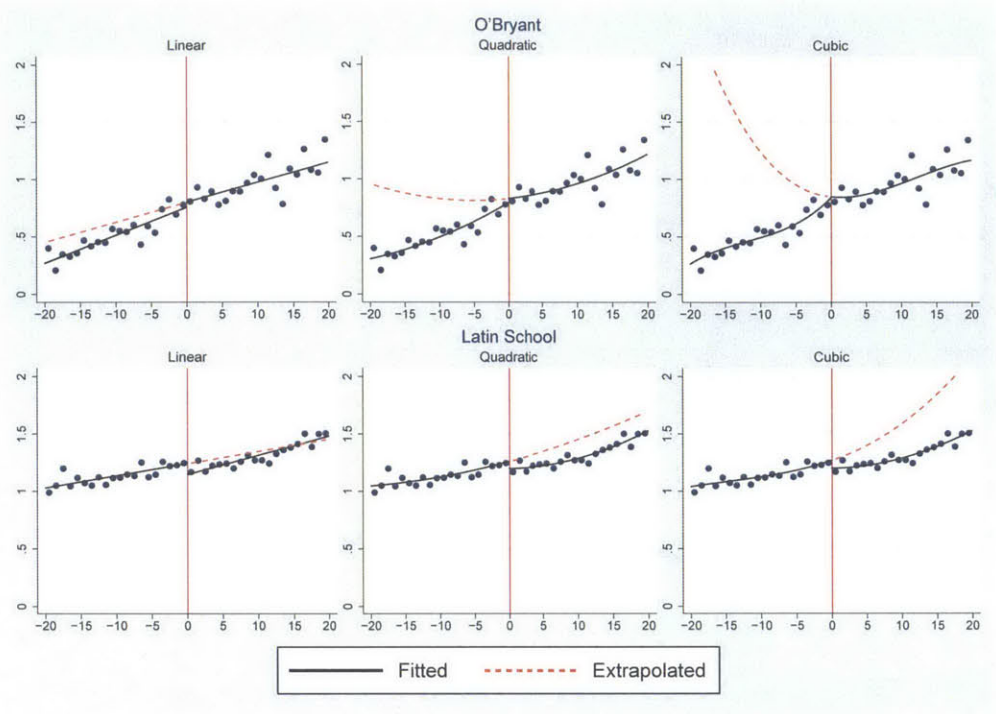
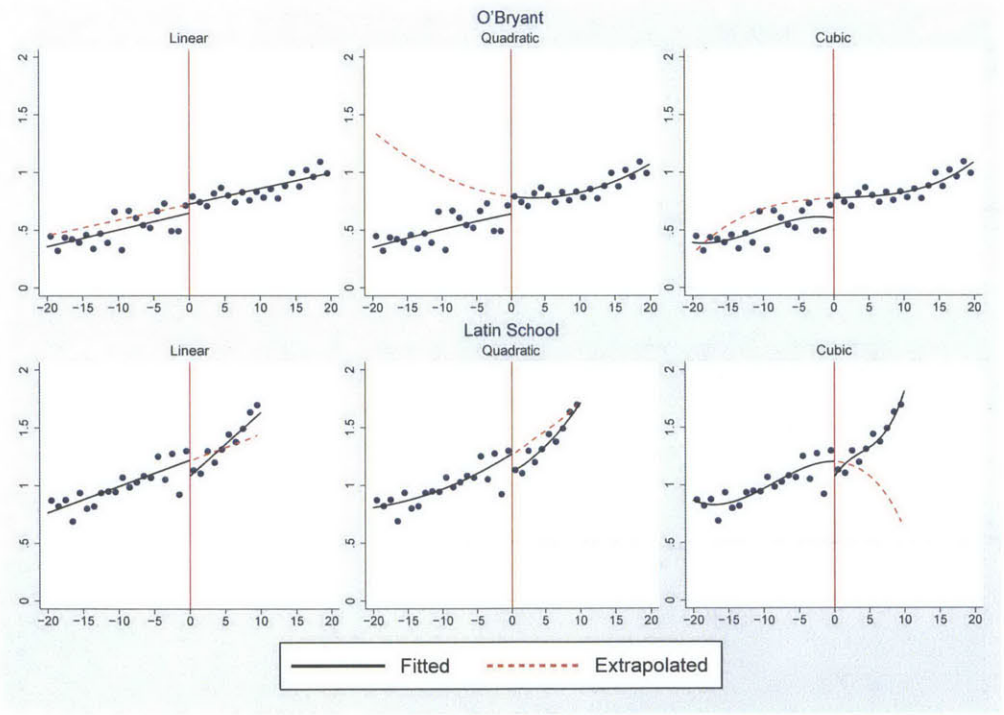


Figure 2-4: Identification of Boston Latin School Effects At and Away from the Cutoff

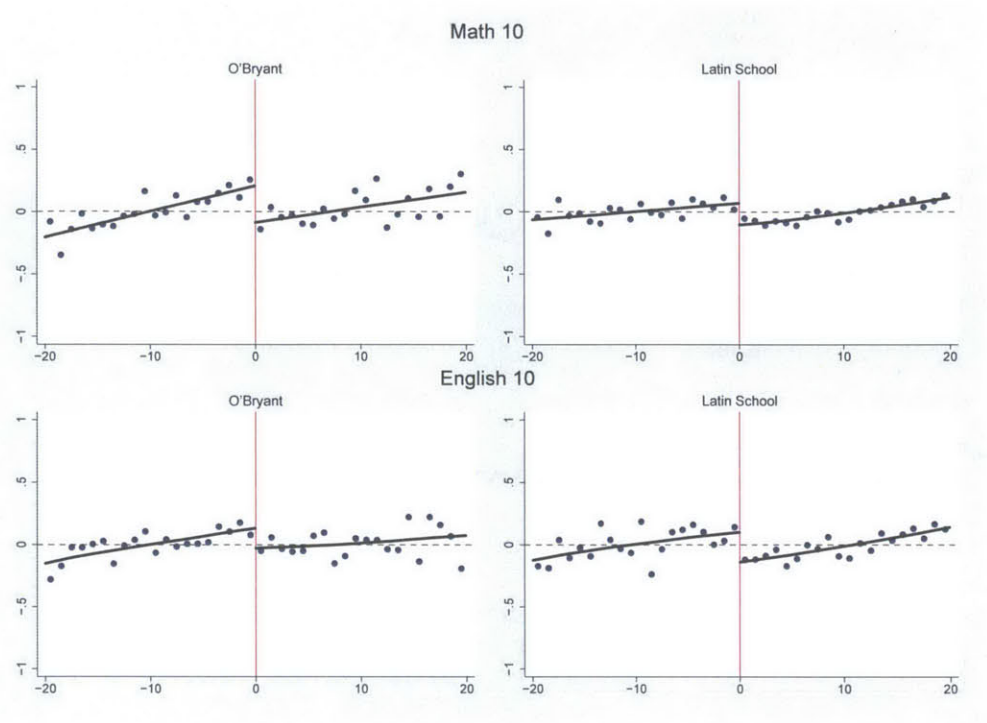


(a) 7th Grade Applicants

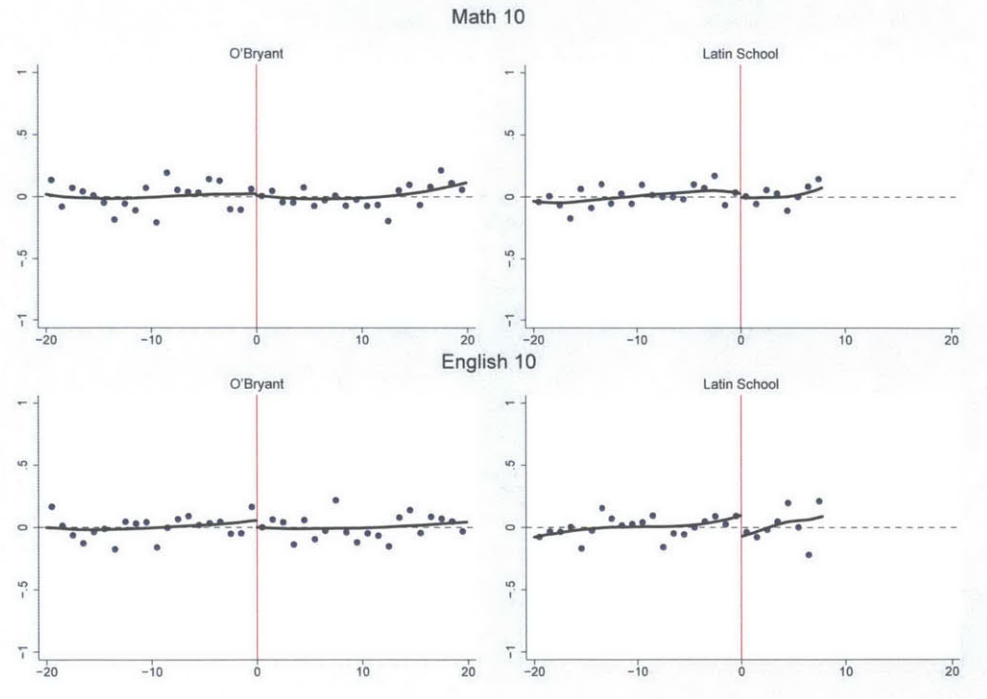


(b) 9th Grade Applicants

Figure 2-5: Parametric Extrapolation at O'Bryant and Boston Latin School for 10th Grade Math



(a) 7th Grade Applicants



(b) 9th Grade Applicants

Figure 2-6: Visual Evaluation of CIA in the Window $[-20, 20]$

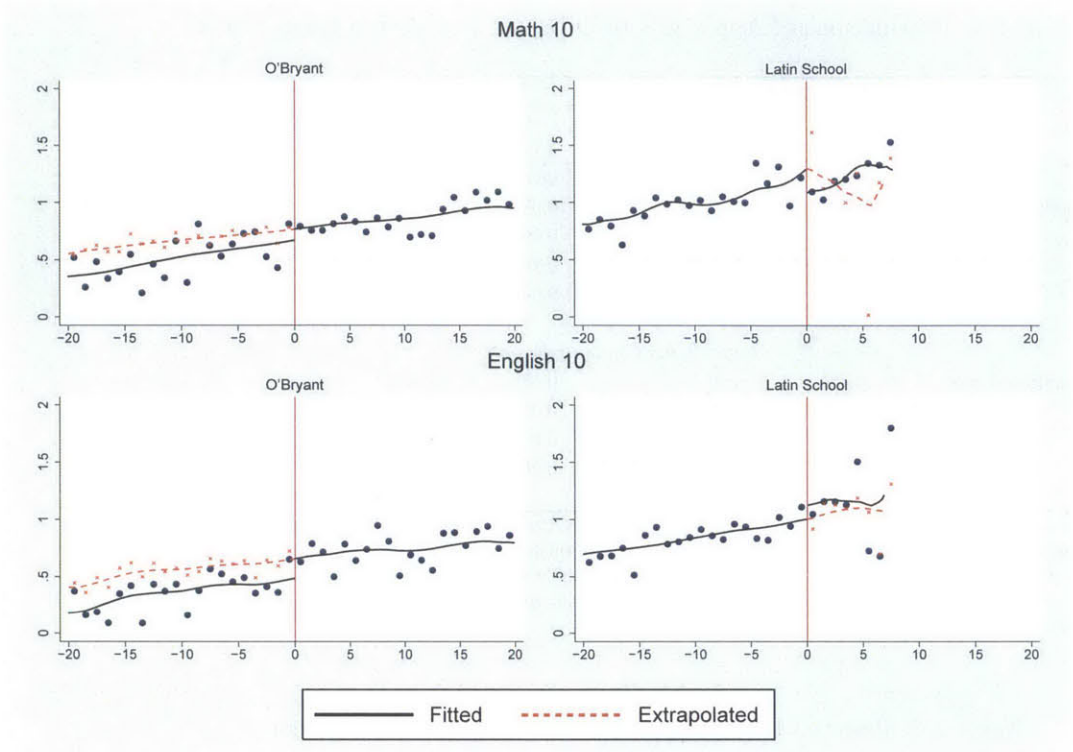


Figure 2-7: CIA-based Estimates of $E[Y_{1i}|r_i = c]$ and $E[Y_{0i}|r_i = c]$ for c in $[-20, 20]$ for 9th Grade Applicants

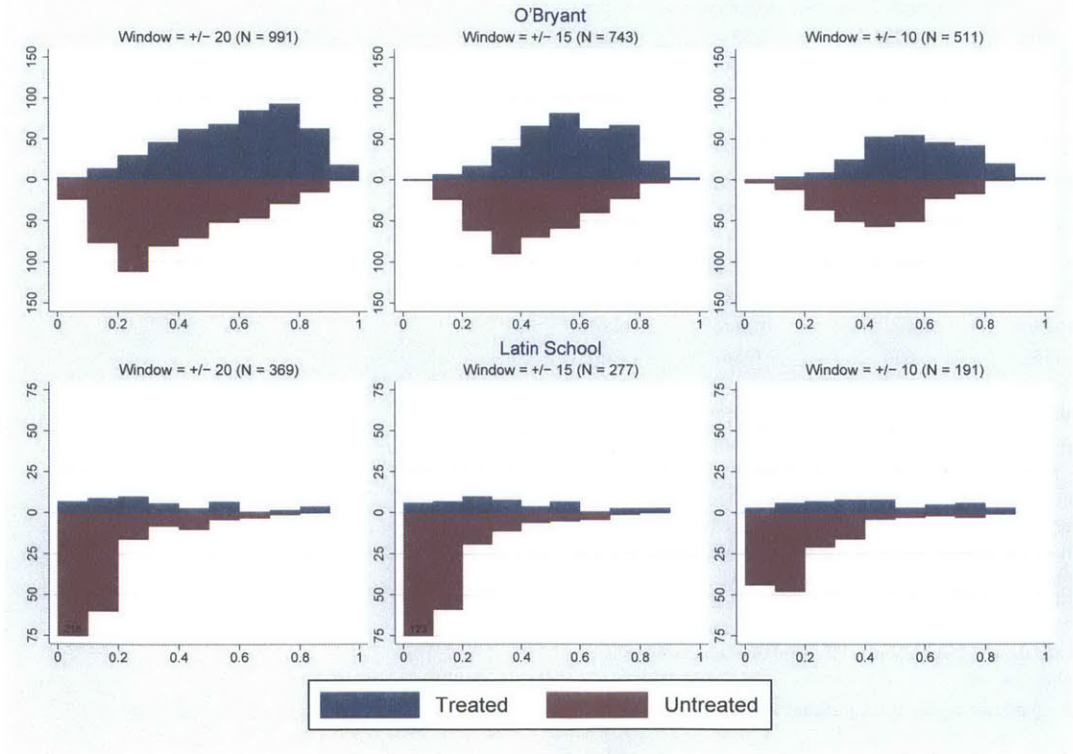


Figure 2-8: Histograms of Estimated Propensity Scores for 9th Grade Applicants to O'Bryant and BLS

Table 2.1: Destinations of Applicants to O'Bryant and Boston Latin School

	O'Bryant		Latin School	
	D=0	D=1	D=0	D=1
	(1)	(2)	(3)	(4)
	Panel A. 7th Grade Applicants			
Traditional Boston public schools	1.00	0.28	0.08	0.05
O'Bryant	0.00	0.72	0.06	0.00
Latin Academy	0.00	0.00	0.86	0.01
Latin School	...	...	0.00	0.93
	Panel B. 9th Grade Applicants			
Traditional Boston public schools	1.00	0.34	0.15	0.04
O'Bryant	0.00	0.66	0.00	0.00
Latin Academy	...	...	0.86	0.02
Latin School	...	...	0.00	0.94

Notes: This table describes the destination schools of Boston exam school applicants. Enrollment rates are measured in the fall admissions cycle following exam school application and estimated using local linear smoothing. The sample of Boston 7th grade applicants includes students who applied for an exam school seat between 1999-2008. The sample of Boston 9th grade applicants includes students who applied for an exam school seat between 2001-2007.

Table 2.2: Reduced Form Estimates for 10th Grade MCAS Scores

	Parametric		Nonparametric	
	O'Bryant	Latin School	O'Bryant	Latin School
	(1)	(2)	(3)	(4)
	Panel A. 7th Grade Applicants			
Math	-0.011	-0.034	0.034	-0.055
	(0.100)	(0.060)	(0.056)	(0.039)
	1832	1854	1699	1467
ELA	0.059	0.021	0.125**	0.000
	(0.103)	(0.095)	(0.059)	(0.061)
	1836	1857	1778	1459
	Panel B. 9th Grade Applicants			
Math	0.166	-0.128	0.128*	-0.144*
	(0.109)	(0.117)	(0.066)	(0.076)
	1559	606	1386	361
ELA	0.191*	0.097	0.180***	0.048
	(0.112)	(0.187)	(0.066)	(0.106)
	1564	607	1532	458

Notes: This table reports estimates of the effects of exam school offers on 10th grade MCAS scores. The sample covers students within 20 standardized units of offer cutoffs. Parametric models include a cubic function of the running variable, allowed to differ on either side of offer cutoffs. Non-parametric estimates use the edge kernel, with bandwidth computed following DesJardins and McCall (2008) and Imbens and Kalyanaraman (2012). Optimal bandwidths were computed separately for each school. Robust standard errors are shown in parentheses. The number of observations is reported below standard errors.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 2.3: Parametric Extrapolation Estimates for 10th Grade Math

	O'Bryant				Latin School			
	c = -1	c = -5	c = -10	c = -15	c = 1	c = 5	c = 10	c = 15
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Panel A: 7th Grade Applicants				
Linear	0.041	0.061	0.085	0.110	-0.076**	-0.051	-0.021	0.010
	(0.052)	(0.057)	(0.072)	(0.093)	(0.035)	(0.040)	(0.049)	(0.061)
	1832	1832	1832	1832	1854	1854	1854	1854
Quadratic	0.063	0.204	0.391*	0.588	-0.056	-0.111	-0.152	-0.161
	(0.075)	(0.125)	(0.237)	(0.384)	(0.051)	(0.088)	(0.162)	(0.261)
	1832	1832	1832	1832	1854	1854	1854	1854
Cubic	0.034	0.167	0.247	0.266	-0.050	-0.096	-0.106	-0.065
	(0.110)	(0.336)	(0.921)	(1.927)	(0.073)	(0.220)	(0.589)	(1.215)
	1832	1832	1832	1832	1854	1854	1854	1854
				Panel B: 9th Grade Applicants				
Linear	0.088	0.083	0.077	0.071	-0.090	0.079	0.291***	0.502***
	(0.057)	(0.059)	(0.070)	(0.088)	(0.065)	(0.063)	(0.108)	(0.168)
	1559	1559	1559	1559	606	606	606	606
Quadratic	0.170**	0.264**	0.427*	0.639*	-0.147*	-0.106	0.078	0.409
	(0.085)	(0.133)	(0.237)	(0.372)	(0.088)	(0.142)	(0.303)	(0.713)
	1559	1559	1559	1559	606	606	606	606
Cubic	0.143	0.069	-0.059	-0.355	-0.061	0.196	0.996	3.094
	(0.119)	(0.327)	(0.851)	(1.735)	(0.118)	(0.338)	(0.910)	(2.543)
	1559	1559	1559	1559	606	606	606	606

Notes: This table reports estimates of effects on 10th grade Math scores away from the RD cutoff at points indicated in the column heading. Columns 1-4 report estimates of the effect of O'Bryant attendance on unqualified O'Bryant applicants. Columns 5-8 report the effects of BLS attendance on qualified BLS applicants. The estimates are based on first, second, and third order polynomials, as indicated in rows of the table. Robust standard errors are shown in parentheses.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 2.4: Conditional Independence Tests

	Math				ELA			
	O'Bryant		Latin School		O'Bryant		Latin School	
	D = 0	D = 1	D = 0	D = 1	D = 0	D = 1	D = 0	D = 1
Window	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
				Panel A. 7th Grade Applicants				
20	0.022***	0.015***	0.008***	0.014***	0.015***	0.006	0.013***	0.018***
	(0.004)	(0.004)	(0.002)	(0.002)	(0.004)	(0.005)	(0.003)	(0.003)
	838	618	706	748	840	621	709	750
15	0.023***	0.015***	0.010***	0.012***	0.014**	0.006	0.007	0.015***
	(0.006)	(0.005)	(0.003)	(0.003)	(0.005)	(0.006)	(0.005)	(0.005)
	638	587	511	517	638	590	514	519
10	0.030***	0.016**	0.010*	0.007	0.024**	0.001	0.012	0.012
	(0.009)	(0.008)	(0.006)	(0.005)	(0.010)	(0.009)	(0.010)	(0.008)
	419	445	335	347	421	447	338	348
				Panel B. 9th Grade Applicants				
20	0.002	0.005	0.008**	0.018	0.003	0.002	0.006	0.055
	(0.004)	(0.003)	(0.003)	(0.028)	(0.004)	(0.004)	(0.005)	(0.053)
	513	486	320	49	516	489	320	50
15	0.010	0.000	0.006	0.018	0.009	-0.000	0.000	0.055
	(0.006)	(0.005)	(0.006)	(0.028)	(0.006)	(0.006)	(0.007)	(0.053)
	375	373	228	49	376	374	229	50
10	0.003	-0.001	0.007	0.018	0.014	-0.004	0.014	0.055
	(0.011)	(0.009)	(0.009)	(0.028)	(0.011)	(0.010)	(0.015)	(0.053)
	253	260	142	49	253	261	142	50

Notes: This table reports regression-based tests of the conditional independence assumption described in the text. Cell entries show the coefficient on the running variable in models for 10th grade math and ELA scores that control for baseline scores, along with indicators for special education status, limited English proficiency, eligibility for free or reduced price lunch, race (black/Asian/Hispanic) and sex, as well as indicators for test year, application year and application preferences. Estimates use only observations to the left or right of the cutoff as indicated in column headings, and were computed in the window width indicated at left. Robust standard errors are reported in parentheses.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 2.5: CIA Estimates of the Effect of Exam School Offers for 9th Grade Applicants

	Linear Reweighting				Propensity Score Weighting			
	Math		ELA		Math		ELA	
	O'Bryant	Latin School	O'Bryant	Latin School	O'Bryant	Latin School	O'Bryant	Latin School
Window	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
20	0.156***	-0.031	0.198***	0.088	0.131***	-0.037	0.236***	0.031
	(0.040)	(0.090)	(0.041)	(0.083)	(0.051)	(0.057)	(0.077)	(0.109)
N untreated	513	320	516	320	509	320	512	320
N treated	486	49	489	50	482	49	485	50
15	0.129***	-0.080	0.181***	0.051	0.103**	-0.070	0.191***	0.003
	(0.044)	(0.055)	(0.044)	(0.093)	(0.052)	(0.054)	(0.062)	(0.111)
N untreated	375	228	376	229	373	228	374	229
N treated	373	49	374	50	370	49	371	50
10	0.091*	-0.065	0.191***	-0.000	0.093*	-0.084	0.166**	-0.062
	(0.054)	(0.059)	(0.053)	(0.097)	(0.054)	(0.062)	(0.068)	(0.133)
N untreated	253	142	253	142	253	142	253	142
N treated	260	49	261	50	258	49	259	50

Notes: This table reports estimates of the effect of exam school offers on MCAS scores for 9th grade applicants to O'Bryant and BLS. Columns 1-4 report results from a linear reweighting estimator, while columns 5-8 report results from inverse propensity score weighting, as described in the text. Controls are the same as used to construct the test statistics except that the propensity score models for Latin School omit test year and application preference dummies. The O'Bryant estimates are effects on nontreated applicants in windows to the left of the admissions cutoff; the BLS estimates are are effects on treated applicants in windows to the right of the cutoff. Standard errors (shown in parentheses) were computed using a nonparametric bootstrap with 500 replications. The table also reports the number of treated and untreated (offered and not offered) observations in each window, in the relevant outcome sample.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 2.6: Fuzzy CIA Estimates of LATE (Exam School Enrollment) for 9th Grade Applicants

	First Stage				LATE			
	Math		ELA		Math		ELA	
	O'Bryant	Latin School	O'Bryant	Latin School	O'Bryant	Latin School	O'Bryant	Latin School
Window	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
20	0.659***	0.898***	0.660***	0.900***	0.225**	-0.031	0.380**	0.060
	(0.062)	(0.054)	(0.062)	(0.052)	(0.088)	(0.217)	(0.183)	(0.231)
N untreated	509	320	512	320	509	320	512	320
N treated	482	49	485	50	482	49	485	50
15	0.666***	0.898***	0.667***	0.900***	0.174**	-0.085	0.302**	0.020
	(0.047)	(0.048)	(0.050)	(0.047)	(0.080)	(0.177)	(0.125)	(0.225)
N untreated	373	228	374	229	373	228	374	229
N treated	370	49	371	50	370	49	371	50
10	0.670***	0.898***	0.678***	0.900***	0.184*	-0.104	0.274**	-0.058
	(0.055)	(0.048)	(0.050)	(0.047)	(0.108)	(0.274)	(0.121)	(0.402)
N untreated	253	142	253	142	253	142	253	142
N treated	258	49	259	50	258	49	259	50

Notes: This table reports fuzzy RD estimates of the effect of exam school enrollment on MCAS scores for 9th grade applicants to O'Bryant and BLS. The O'Bryant estimates are effects on nontreated applicants in windows to the left of the admissions cutoff; the BLS estimates are for treated applicants in windows to the right of the cutoff. The first stage estimates in columns 1-4 and the estimated causal effects in columns 5-8 are from a modified propensity-score style weighting estimator described in the text. Standard errors (shown in parentheses) were computed using a nonparametric bootstrap with 500 replications. The table also reports the number of treated and untreated (offered and not offered) observations in each window, in the relevant outcome sample.

* significant at 10%; ** significant at 5%; *** significant at 1%

Table 2.7: Fuzzy CIA Estimates of Average Causal Response (Years of Exam School Enrollment) for 9th Grade Applicants

	First Stage				ACR			
	Math		ELA		Math		ELA	
	O'Bryant	Latin School	O'Bryant	Latin School	O'Bryant	Latin School	O'Bryant	Latin School
Window	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
20	1.394***	1.816***	1.398***	1.820***	0.112***	-0.017	0.142***	0.048
	(0.064)	(0.096)	(0.065)	(0.093)	(0.029)	(0.050)	(0.030)	(0.045)
N untreated	513	320	516	320	513	320	516	320
N treated	486	49	489	50	486	49	489	50
15	1.359***	1.816***	1.363***	1.820***	0.095***	-0.044	0.133***	0.028
	(0.064)	(0.099)	(0.064)	(0.089)	(0.032)	(0.031)	(0.034)	(0.047)
N untreated	375	228	376	229	375	228	376	229
N treated	373	49	374	50	373	49	374	50
10	1.320***	1.816***	1.312***	1.820***	0.069	-0.036	0.145***	-0.000
	(0.080)	(0.095)	(0.080)	(0.089)	(0.043)	(0.031)	(0.041)	(0.054)
N untreated	253	142	253	142	253	142	253	142
N treated	260	49	261	50	260	49	261	50

Notes: This table reports fuzzy RD estimates of the effect of years of exam school enrollment on MCAS scores for 9th grade applicants to O'Bryant and BLS. The O'Bryant estimates are effects on nontreated applicants in windows to the left of the admissions cutoff; the BLS estimates are for treated applicants in windows to the right of the cutoff. The first stage estimates in columns 1-4 and the estimated causal effects in columns 5-8 are from a modified linear 2SLS estimator described in the text. Standard errors (shown in parentheses) were computed using a nonparametric bootstrap with 500 replications. The table also reports the number of treated and untreated (offered and not offered) observations in each window, in the relevant outcome sample.

* significant at 10%; ** significant at 5%; *** significant at 1%

2.8 Appendix

Defining Sharp Samples

Boston exam school applicants rank up to three schools in order of preference, while schools rank their applicants according to an average of GPA and ISEE scores. Applicants are ranked only for schools to which they've applied, so applicants with the same GPA and ISEE scores might be ranked differently at different schools depending on where they fall in each school's applicant pool (each also school weights ISEE and GPA a little differently). Applicants are ranked at every school to which they apply, regardless of how they've ordered schools. Student-proposing deferred acceptance (DA) generates offers from student preference and school-specific rankings as follows:

- In round 1: Each student applies to his first choice school. Each school rejects the lowest-ranked applicants in excess of capacity, with the rest provisionally admitted (students not rejected at this step may be rejected in later steps.)
- At round $\ell > 1$: Students rejected in Round ℓ-1 apply to their next most preferred school (if any). Each school considers these students *and* provisionally admitted students from the previous round, rejecting the lowest-ranked applicants in excess of capacity from this combined pool and producing a new provisional admit list (again, students not rejected at this step may be rejected in later steps.)

The DA algorithm terminates when either every student is matched to a school or every unmatched student has been rejected by every school he has ranked.

Let τ_k denote the rank of the last applicant offered a seat at school k; let c_{ik} denote student i's composite score at school k; and write the vector of composite scores as $\mathbf{c}_i = (c_{i1}, c_{i2}, c_{i3})$, where c_{ik} is missing if student i did not rank school k. A dummy variable $q_i(k) = 1[c_{ik} \leq \tau_k]$ indicates that student i qualified for school k by clearing τ_k (rank and qualification at k are missing for applicants who did not rank k). Finally, let p_{ik} denote student i's kth choice and represent i's preference list by $\mathbf{p}_i = (p_{i1}, p_{i2}, p_{i3})$, where $p_{ik} = 0$ if the list is incomplete. Students who ranked and qualified for a school will not be offered a seat at that school if they get an offer from a more preferred school. With three schools ranked, applicant i is offered a seat at school k in one of three ways:

- The applicant ranks school k first and qualifies: $(\{p_{i1} = k\} \cap \{q_i(k) = 1\})$.
- The applicant doesn't qualify for his first choice, ranks school k second and qualifies there: $(\{q_i(p_{i1}) = 0\} \cap \{p_{i2} = k\} \cap \{q_i(k) = 1\})$.
- The applicant doesn't qualify at his top two choices, ranks school k third, and qualifies there: $(\{q_i(p_{i1}) = q_i(p_{i2}) = 0\} \cap \{p_{i3} = k\} \cap \{q_i(k) = 1\})$.

We summarize the relationship between composite scores, cutoffs, and offers by letting O_i be student i's offer, with the convention that $O_i = 0$ means no offer. DA determines O_i as follows:

$$O_i = \sum_{j=1}^{J} p_{ij} q_i(p_{ij}) \left[\prod_{\ell=1}^{j-1} (1 - q_i(p_{i\ell})) \right].$$

The formulation shows that the sample for which offers at school k are deterministically linked with the school-k composite score - the *sharp sample* for school k - is the union of three sets of applicants:

- Applicants who rank k first, so $(p_{i1} = k)$
- Applicants unqualified for their first choice, ranking k second, so $(q_i(p_{i1}) = 0 \cap p_{i2} = k)$
- Applicants unqualified for their top two choices, ranking k third, so $((q_i(p_{i1}) = q_i(p_{i2}) = 0) \cap p_{i3} = k)$.

All applicants are in at least one sharp sample (at the exam school they rank first), but can be in more than one. For example, a student who ranked BLS first, but did not qualify there, is also in the sharp sample for BLA if he ranked BLA second.

A possible concern with nonparametric identification strategies using sharp samples arises from the fact that the sharp sample itself may change discontinuously at the cutoff. Suppose, for example, that two schools have the same cutoff and a common running variable. Some students rank school 2 ahead of school 1 and some rank school 1 ahead of school 2. The sharp sample for school 1 includes both those who rank 1 first and those who rank 2 first but are disqualified there. This second group appears only to the left of the common cutoff, changing the composition of the sharp sample for school 1 (with a similar argument applying to the sharp sample for school 2). In view of this possibility, all estimating equations include dummies for applicants' preference orderings over schools.

Proof of Theorem 1

We continue to assume that GCIA and other LATE assumptions hold. Given these assumptions, Theorem 3.1 in Abadie (2003) implies that for any measurable function, $g(y_i, W_i, x_i)$, we have

$$E[g(y_i, W_i, x_i) \mid x_i, W_{1i} > W_{0i}] = \frac{1}{P[W_{1i} > W_{0i} \mid x_i]} E[\kappa(W_i, D_i, x_i)\, g(y_i, W_i, x_i) \mid x_i] \quad (2.27)$$

where

$$\kappa(W_i, D_i, x_i) = 1 - \frac{W_i(1 - D_i)}{1 - P[D_i = 1 \mid x_i]} - \frac{(1 - W_i) D_i}{P[D_i = 1 \mid x_i]}$$

and

$$E[g(Y_{Wi}, x_i) \mid x_i, W_{1i} > W_{0i}] = \frac{1}{P[W_{1i} > W_{0i} \mid x_i]} E[\kappa_W(W_i, D_i, x_i)\, g(y_i, x_i) \mid x_i],$$

where $W \in \{0,1\}$ and

$$\begin{aligned}
\kappa_0(W_i, D_i, x_i) &= (1 - W_i) \frac{P[D_i = 1 \mid x_i] - D_i}{(1 - P[D_i = 1 \mid x_i]) P[D_i = 1 \mid x_i]} \\
\kappa_1(W_i, D_i, x_i) &= W_i \frac{D_i - P[D_i = 1 \mid x_i]}{(1 - P[D_i = 1 \mid x_i]) P[D_i = 1 \mid x_i]}.
\end{aligned}$$

Using the GCIA, we can simplify as follows:

$$\begin{aligned}
&E[g(Y_{Wi}, x_i) \mid W_{1i} > W_{0i}, 0 < r_i \leq c] \\
&= E\{E[g(Y_{Wi}, x_i) \mid x_i, W_{1i} > W_{0i}] \mid W_{1i} > W_{0i}, 0 < r_i \leq c\} \\
&= \int \frac{1}{P[W_{1i} > W_{0i} \mid x_i]} E[\kappa_W(W_i, D_i, x_i) g(y_i, x_i) \mid X] dP[x_i \mid W_{1i} > W_{0i}, 0 < r_i \leq c] \\
&= \frac{1}{P[W_{1i} > W_{0i} \mid 0 < r_i \leq c]} \int E[\kappa_W(W_i, D_i, x_i) g(y, x_i) \mid x_i] \frac{P[0 < r_i \leq c \mid x_i]}{P[0 < r_i \leq c]} dP[x_i] \quad (2.28) \\
&= \frac{1}{P[W_{1i} > W_{0i} \mid 0 < r_i \leq c]} E\left[\kappa_W(W_i, D_i, x_i) \frac{P[0 < r_i \leq c \mid x_i]}{P[0 < r_i \leq c]} g(y_i, x_i)\right].
\end{aligned}$$

This implies that LATE can be written:

$$\begin{aligned}
&E[Y_{1i} - Y_{0i} \mid W_{1i} > W_{0i}, 0 < r_i \leq c] \\
&= E[Y_{1i} \mid W_{1i} > W_{0i}, 0 < r_i \leq c] - E[Y_{0i} \mid W_{1i} > W_{0i}, 0 < r_i \leq c] \\
&= \frac{1}{P[W_{1i} > W_{0i} \mid 0 < r_i \leq c]} E\left[\psi(D_i, x_i) \frac{P[0 < r_i \leq c \mid x_i]}{P[0 < r_i \leq c]} y_i\right]
\end{aligned}$$

where

$$\begin{aligned}
\psi(D_i, x_i) &= \kappa_1(W_i, D_i, x_i) - \kappa_0(W_i, D_i, x_i) \\
&= \frac{D_i - P[D_i = 1 \mid x_i]}{(1 - P[D_i = 1 \mid x_i]) P[D_i = 1 \mid x_i]}.
\end{aligned}$$

Finally, by setting $g(y_i, W_i, x_i) = 1$ in equation (2.27) we get:

$$P[W_{1i} > W_{0i} \mid x_i] = E[\kappa(W_i, D_i, x_i) \mid x_i].$$

Using the same steps as in equation (2.28), the GCIA implies:

$$\begin{aligned}
P[W_{1i} > W_{0i} \mid 0 < r_i \leq c] &= E\{P[W_{1i} > W_{0i} \mid x_i] \mid 0 < r_i \leq c\} \\
&= E\left[\kappa(W_i, D_i, x_i) \frac{P[0 < r_i \leq c \mid x_i]}{P[0 < x_i \leq c]}\right].
\end{aligned}$$

Proof of Theorem 2

Theorem 1 in Angrist and Imbens (1995) implies:

$$\begin{aligned} E[y_i \mid D_i = 1, x_i] - E[y_i \mid D_i = 0, x_i] &= \sum_j P[w_{1i} \geq j > w_{0i} \mid x_i] \, E[Y_{ji} - Y_{j-1,i} \mid w_{1i} \geq j > w_{0i}, x_i] \\ E[w_i \mid D_i = 1, x_i] - E[w_i \mid D_i = 0, x_i] &= \sum_j P[w_{1i} \geq j > w_{0i} \mid x_i] \, . \end{aligned}$$

Given the GCIA, we have:

$$\begin{aligned} & E\{E[y_i \mid D_i = 1, x_i] - E[y_i \mid D_i = 0, x_i] \mid 0 < r_i \leq c\} \\ &= \sum_j \int P[w_{1i} \geq j > w_{0i} \mid x_i] \, E[Y_{ji} - Y_{j-1,i} \mid w_{1i} \geq j > w_{0i}, x_i] \, dP[x_i \mid 0 < r_i \leq c] \\ &= \sum_j \int P[w_{1i} \geq j > w_{0i} \mid x_i, 0 \leq r_i \leq c] \, E[Y_{ji} - Y_{j-1,i} \mid w_{1i} \geq j > w_{0i}, x_i] \, dP[x_i \mid 0 < r_i \leq c] \\ &= \sum_j P[w_{1i} \geq j > w_{0i} \mid 0 < r_i \leq c] \\ &\quad \times \int E[Y_{ji} - Y_{j-1,i} \mid w_{1i} \geq j > w_{0i}, x_i] \, dP[x_i \mid w_{1i} \geq j > w_{0i}, 0 < r_i \leq c] \\ &= \sum_j P[w_{1i} \geq j > w_{0i} \mid 0 < r_i \leq c] \, E[Y_{ji} - Y_{j-1,i} \mid w_{1i} \geq j > w_{0i}, 0 < r_i \leq c] \, . \end{aligned}$$

The GCIA can similarly be shown to imply:

$$\begin{aligned} & E[E[w_i \mid D_i = 1, x_i] - E[w_i \mid D_i = 0, x_i] \mid 0 < r_i \leq c] \\ &= \sum_j P[w_{1i} \geq j > w_{0i} \mid 0 < r_i \leq c] \, . \end{aligned}$$

Combining these results, the ACR can be written:

$$\begin{aligned} & \frac{E\{E[y_i \mid D_i = 1, x_i] - E[y_i \mid D_i = 0, x_i] \mid 0 < r_i \leq c\}}{E\{E[w_i \mid D_i = 1, x_i] - E[w_i \mid D_i = 0, x_i] \mid 0 < r_i \leq c\}} \\ &= \sum_j \nu_{jc} E[Y_{ji} - Y_{j-1,i} \mid w_{1i} \geq j > w_{0i}, 0 < r_i \leq c] \end{aligned}$$

where

$$\nu_{ijc} = \frac{P[w_{1i} \geq j > w_{0i} \mid 0 < r_i \leq c]}{\sum_\ell P[w_{1i} \geq \ell > w_{0i} \mid 0 < r_i \leq c]} .$$

Chapter 3

Adaptive Bandwidth Choice for the Regression Discontinuity Design

3.1 Introduction

The regression discontinuity (RD) design, originating from Thistlewhite and Campbell (1960), has become a popular approach in economics to identifying causal effects of various treatments. In this design the treatment of interest is either fully or partly determined by whether the value of an observed covariate, often referred to as the running variable, lies below or above a known cutoff. Under relatively weak assumptions, this allows one to identify the causal effect of the treatment for individuals at the cutoff. RD designs have been used to study, for instance, the effect of class size on student achievement (Angrist and Lavy, 1999), parental valuation of school quality (Black, 1999), the effect of financial aid on college enrollment (van der Klaauw, 2002), and the effect of Head Start on child mortality (Ludwig and Miller, 2007). In addition, Hahn, Todd, and van der Klaauw (2001) and Porter (2003), among others, have made important contributions to the literature on identification and estimation of treatment effects in the RD design.[1]

The consistency of the RD estimator relies heavily on the researcher's ability to correctly specify the functional form for the relationship between the running variable and the outcome and the relationship between the running variable and the treatment. This has lead to a widespread interest in nonparametric approaches to estimating these relationships. A common approach in the recent literature has been to use local polynomial regression (LPR), especially local linear regression. As the performance of LPR-based methods depends heavily on the choice of a smoothing parameter, often referred to as the bandwidth, a key question in implementing these methods is how to choose this parameter.

Traditionally, the bandwidth choice in empirical work using LPR-based RD estimators has been based on

[1] For extensive surveys of the literature, see Cook (2008), Imbens and Lemieux (2008), van der Klaauw (2008), and Lee and Lemieux (2010).

either ad hoc procedures or on approaches that are not directly suited to the RD design (Ludwig and Miller, 2005; DesJardins and McCall, 2008). However, in a recent influential paper Imbens and Kalyanaraman (2012) studied in depth the problem of optimal bandwidth choice for local linear regression-based RD estimator and proposed an algorithm that can be used to obtain a consistent estimator of the asymptotically optimal RD bandwidth.[2]

This paper contributes to the literature by proposing an adaptive bandwidth choice algorithm for the LPR-based RD estimator by building on previous work by Schucany (1995) and Gerard and Schucany (1997).[3] The algorithm is adaptive in the sense that it allows for different choices for the order of polynomial and kernel function. In addition, the algorithm automatically takes into account the inclusion of additional covariates as well as alternative assumptions on the variance-covariance structure of the error terms. Thus, the proposed algorithm provides a convenient approach to bandwidth choice that retains its validity in various settings.

I show that the proposed algorithm produces a consistent estimator of the asymptotically optimal bandwidth. Furthermore, the resulting RD estimator satisfies the asymptotic optimality criterion of Li (1987) and converges to the true parameter value at the optimal nonparametric rate (Stone, 1982; Porter, 2003). I also provide Monte Carlo evidence illustrating that that the propoced algorithm works well in finite sample and compares favorably to the algorithm by Imbens and Kalyanaraman (2012).

The rest of the paper is structured as follows. Section 2 reviews the RD design and the LPR-based RD estimator. Section 3 introduces the proposed bandwidth choice algorithm and discusses its asymptotic properties. Section 4 presents Monte Carlo evidence illustrating the finite-sample performance of the proposed algorithm. Section 5 concludes.

3.2 Regression Discontinuity Design

3.2.1 Setting and Parameter of Interest

Suppose one is interestein the causal effect of a binary treatment on some outcome. Let D denote an indicator that equals 1 if an individual receives the treatment and 0 otherwise. Furthermore, let Y_1 and Y_0 denote the potential outcomes when an individual receives and does not receive the treatment. The observed outcome of an individual, denoted by Y, is

$$Y = (1 - D) \times Y_0 + D \times Y_1.$$

In a sharp regression discontinuity (SRD) design D is a deterministic function of a continuous running

[2]See also Arai and Ichimura (2013) for an alternative approach to optimal bandwidth choice for the local liner regression-based RD estimator. Furthermore, Calonico, Cattaneo, and Titiunik (2014) discuss nonparametric estimation of robust confidence intervals for the local liner regression-based RD estimator.

[3]Similar approaches to optimal bandwidth choice have also been proposed by Ruppert (1997), Doksum, Peterson, and Samarov (2000), and Prewitt (2003).

variable R:[4]

$$D = 1(R \geq c)$$

where $1(\cdot)$ is an indicator function equal to 1 if the statement in parentheses is true and 0 otherwise. In words, all individuals with the value of R at or above a cutoff c are assigned to the treatment group while all individuals with the value of R below the cutoff c are assigned to the control group. Furthermore, there is perfect compliance with the treatment assignment: all of the individuals assigned to the treatment group receive the treatment whereas none of the individuals assigned to the control group receive the treatment.

Given the treatment assignment mechanism, a natural parameter of interest in the SRD design is

$$\tau = E[Y_1 - Y_0 \mid R = c],$$

that is, the average effect of the treatment for individuals at the cutoff. Suppose that $E[Y_1 \mid R = r]$ and $E[Y_0 \mid R = r]$ exist and are continuous at $R = c$. Then

$$\tau = \lim_{r \downarrow c} E[Y \mid R = r] - \lim_{r \uparrow c} E[Y \mid R = r]$$

where $m(r) = E[Y \mid R = r]$, $m_+(c) = m(r)$ and $m_-(c) = m(r)$. Thus, under relatively mild assumptions τ is nonparametrically identified and given by the difference in the limits of two conditional expectation functions at the cutoff c.

3.2.2 Estimation using Local Polynomial Regression

I focus in this paper on the estimation of τ using separate LPRs on both sides of the cutoff.[5] An attractive property of the LPR-based approach is that it allows one to obtain a consistent estimator of τ without reliance on strong functional form assumptions. Moreover, the LPR-based approach reduces (and under some assumptions even eliminates) the bias that afflicts other nonparametric regression function estimates at boundary points.

Suppose we observe a sample (Y_i, R_i), $i = 1, \ldots, n$. The LPR-based estimator of τ using a polynomial of order p , kernel $K(u)$, and bandwidth h is given by

$$\hat{\tau}_p(h) = \hat{\alpha}_p^+(h) - \hat{\alpha}_p^-(h)$$

[4] I focus solely on SRD design in this paper. Fuzzy RD design is a straightforward extension that I leave for future research.
[5] For a comprehensive treatment of LPR methods, see Fan and Gijbels (1996).

where

$$\begin{bmatrix} \hat{\alpha}_p^+(h) \\ \hat{\beta}_{1,p}^+(h) \\ \vdots \\ \hat{\beta}_{p,p}^+(h) \end{bmatrix} = \underset{\alpha,\{\beta_k\}_{k=1}^p}{\arg\min} \sum_{i=1}^n 1(R_i \geq c) K\left(\frac{R_i - c}{h}\right) \left(Y_i - \alpha - \sum_{k=1}^p \beta_k (R_i - c)^k\right)^2$$

and

$$\begin{bmatrix} \hat{\alpha}_p^-(h) \\ \hat{\beta}_{1,p}^-(h) \\ \vdots \\ \hat{\beta}_{p,p}^-(h) \end{bmatrix} = \underset{\alpha,\{\beta_k\}_{k=1}^p}{\arg\min} \sum_{i=1}^n 1(R_i < c) K\left(\frac{R_i - c}{h}\right) \left(Y_i - \alpha - \sum_{k=1}^p \beta_k (R_i - c)^k\right)^2.$$

I have written the estimator $\hat{\tau}_p(h)$ in a way that makes explicit its dependence on the choice of the order of polynomial p and the bandwidth h. The estimator $\hat{\tau}_p(h)$ also depends on the choice of the kernel $K(u)$, but this does not play a key role in what follows. Covariates could easily be included in the model, but I abstract away from this for notational simplicity.

Let

$$Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix}, \; X_p = \begin{bmatrix} 1 & (R_1 - c) & \cdots (R_1 - c)^p \\ \vdots & \vdots & \vdots \\ 1 & (R_n - c) & \cdots (R_n - c)^p \end{bmatrix}$$

and define

$$\begin{aligned} \hat{\beta}_p^+(h) &= \left(X_p' W_h^+ X_p'\right)^{-1} X_p' W_h^+ Y \\ \hat{\beta}_p^-(h) &= \left(X_p' W_h^- X_p'\right)^{-1} X_p' W_h^- Y \end{aligned}$$

where

$$\begin{aligned} W_h^+ &= diag\left[1(R_i \geq c) K\left(\frac{R_i - c}{h}\right)\right] \\ W_h^- &= diag\left[1(R_i < c) K\left(\frac{R_i - c}{h}\right)\right]. \end{aligned}$$

We can now write $\hat{\tau}_p(h)$ equivalently as

$$\hat{\tau}_p(h) = e_1' \left(\hat{\beta}_p^+(h) - \hat{\beta}_p^-(h)\right)$$

where e_1 is a $(p+1) \times 1$ vector with one as its first element and zeros as the other elements.

Using standard results for Weighted Least Squares (WLS) estimators one can write the heteroskedasticity-robust variance estimators for $\hat{\beta}_p^+(h)$ and $\hat{\beta}_p^-(h)$ as

$$\begin{aligned}
\hat{v}_p^+(h) &= \left(X_p' W_h^+ X_p'\right)^{-1} X_p' W_h^+ \hat{\Sigma}_{p,h}^+ W_h^+ X_p \left(X_p' W_h^+ X_p'\right)^{-1} \\
\hat{v}_p^-(h) &= \left(X_p' W_h^- X_p'\right)^{-1} X_p' W_h^- \hat{\Sigma}_{p,h}^- W_h^+ X_p \left(X_p' W_h^- X_p'\right)^{-1}
\end{aligned}$$

where

$$\begin{aligned}
\hat{\Sigma}_{p,h}^+ &= diag\left[\left(Y_i - X_{ip}' \hat{\beta}_p^+(h)\right)^2\right] \\
\hat{\Sigma}_{p,h}^- &= diag\left[\left(Y_i - X_{ip}' \hat{\beta}_p^-(h)\right)^2\right]
\end{aligned}$$

and X_{pi}' is the i^{th} row of X_p. Thus,

$$\hat{v}_p(h) = e_1' \left(\hat{v}_p^+(h) + \hat{v}_p^-(h)\right) e_1$$

provides a heteroskedasticity-robust variance estimator for $\hat{\tau}_p(h)$.[6]

As was mentioned above, there are in general three decisions one has to make when implementing LPR-based estimators: order of polynomial p, kernel $K(u)$ and bandwidth h. I focus in this paper on the choice of h conditional on the choices of p and $K(u)$. This is motivated by the observation that bandwidth choice is commonly viewed as the key decision when implementing LPR-based estimators. As the bandwidth choice algorithm proposed in this paper applies to generic p and $K(u)$, I will make only some remarks regarding these choices.

A common approach in the empirical literature is to use local linear regression-based estimators. These are convenient in practice as the number of parameters needed to be estimated is relativelely small. These estimators have also been shown to have attractive bias properties at boundary points (Fan and Gijbels, 1992) and to obtain the optimal convergence rate (Stone, 1982; Porter, 2003).

Common choices for the kernel include the uniform kernel $K(u) = \frac{1}{2} 1(|u| \leq 1)$, the Epanechnikov kernel $K(u) = \frac{3}{4}\left(1 - u^2\right) 1(|u| \leq 1)$ and the triangular kernel (some authors refer to this as the edge kernel) $K(u) = (1 - |u|) 1(|u| \leq 1)$. The popularity of the uniform kernel is mainly due to its practical convenience while the Epanechnikov kernel has been shown to be optimal for estimation problems at interior points (Fan, Gasser, Gijbels, Brockmann, and Engel, 1997). The triangular kernel is instead the most appropriate choice for the RD design as it has been shown to be optimal for estimation problems at boundary points (Cheng, Fan, and Marron, 1997). Imbens and Kalyanaraman (2012), for instance, focus on this kernel in

[6]Note that in practice one can compute $\hat{\tau}_p(h)$ and $\hat{v}_p(h)$ using WLS with full set of interactions for the running variable controls and the indicator variable for the value of the running variable being above the cutoff. However, for simplicity I use the above notation throughout the paper.

their bandwidth choice algorithm.

3.3 Optimal Bandwidth Choice

3.3.1 Infeasible Bandwidth Choice

The optimality criteria I use in this paper is the Mean Squared Error (MSE) which can be written as

$$\begin{aligned} MSE\left[\hat{\tau}_p(h)\right] &= E\left[\left(\hat{\tau}_p(h)-\tau\right)^2\right] \\ &= E\left[\hat{\tau}_p(h)-\tau\right]^2 + E\left[\left(\hat{\tau}_p(h)-E\left[\hat{\tau}_p(h)\right]\right)^2\right]. \end{aligned}$$

In words, the MSE equals the sum of the squared bias and the variance of $\hat{\tau}_p(h)$. While the estimation of the variance of $\hat{\tau}_p(h)$ is relatively straightforward, the estimation of the bias of $\hat{\tau}_p(h)$ is problematic. Thus, it is typically difficult to obtain a good estimator of the bandwidth that minimizes the MSE.[7] I follow instead the standard approach in the literature on LPR-based methods and focus on the first-order asymptotic approximation of the MSE refered to as the Asymptotic Mean Squared Error (AMSE). Furthermore, I focus on the case in which the bandwidth is restricted to be the same on both sides of the cutoff as opposed to choosing a different bandwidth to the left and to the right of the cutoff.[8]

I will next state the assumptions used throughout this paper.

Assumption J.

1. The observations (Y_i, R_i), $i = 1, \ldots, n$, *are independent and identically distributed.*

2. The conditional expectation function $m(r) = E[Y_i \mid R_i = r]$ *is at least* $p+2$ *times continuously differentiable at* $r \neq c$. *Let* $m^{(k)}(r)$ *denote the* k^{th} *derivative of* $m(r)$. $\left|m^{(k)}(r)\right|$, $k = 0, \ldots, p+2$, *are uniformly bounded on* $(c, c+M]$ *and* $[c-M, c)$ *for some* $M > 0$. $\left|m_-^{(k)}(c)\right|$ *and* $\left|m_+^{(k)}(c)\right|$, $k = 0, \ldots, p+2$, *exist and are finite, where* $m_-^{(k)}(c)$ *and* $m_+^{(k)}(c)$ *denote the left and right limit of* $m^{(k)}(r)$ *at the cutoff* c.

3. The marginal distribution of the running variable R_i, *denoted by* $f(r)$, *is continuous, bounded and bounded away from zero around* c.

4. Let ϵ_i *denote the residual* $Y_i - m(R_i)$. *The conditional variance function* $\sigma^2(r) = Var[\epsilon_i \mid R_i = r]$ *is uniformly bounded on* $(c, c+M]$ *and* $[c-M, c)$ *for some* $M > 0$. *The left and right limits of* $\sigma^2(r)$ *at the cutoff, denoted by* $\sigma_+^2(c)$ *and* $\sigma_-^2(c)$, *exist and are finite.*

5. $E\left[|\epsilon_i|^4 \mid R_i = r\right]$ *is uniformly bounded on* $(c, c+M]$ *and* $[c-M, c)$ *for some* $M > 0$. *The limits* $\lim_{r \downarrow c} E\left[|\epsilon_i|^4 \mid R_i = r\right]$ *and* $\lim_{r \uparrow c} E\left[|\epsilon_i|^4 \mid R_i = r\right]$ *exist and are finite.*

6. The kernel $K(u)$ *is non-negative, bounded, different from zero on the compact interval* $[0, 1]$ *and continuous on the open interval* $(0, 1)$.

[7] See also the discussion in Imbens and Kalyanaraman (2012).

[8] Arai and Ichimura (2013) propose a bandwidth choice algorithm for the local linear regression-based estimator that uses separate bandwidths to the left and right of the cutoff.

We can now formally define the AMSE of $\hat{\tau}_p(h)$ as

$$AMSE[\hat{\tau}_p(h)] = B_p^2 h^{2(p+1)} + \frac{V_p}{nh}$$

where

$$\begin{aligned} B_p &= \frac{m_+^{(p+1)}(c)}{(p+1)!} e_1' \Gamma_+^{-1} \delta_+ - \frac{m_-^{(p+1)}(c)}{(p+1)!} e_1' \Gamma_-^{-1} \delta_- \\ V_p &= \frac{\sigma_+^2(c)}{f(c)} \Gamma_+^{-1} \Lambda_+ \Gamma_+^{-1} + \frac{\sigma_-^2(c)}{f(c)} \Gamma_-^{-1} \Lambda_- \Gamma_-^{-1}. \end{aligned}$$

The vectors/matrices Γ_+, Γ_-, δ_+, δ_-, Λ_+, and Λ_-, defined in the Appendix, depend only on $K(u)$. The AMSE provides an approximation to the MSE for small h and large nh, as shown in Theorem 1. The first term of the AMSE corresponds to the square of the leading term of an asymptotic approximation of the bias of $\hat{\tau}_p(h)$. The second term of the AMSE corresponds to the leading term of an asymptotic approximation of the variance of $\hat{\tau}_p(h)$. The expression illustrates the bias-variance tradeoff inherent in the problem of choosing h: using a larger bandwidth reduces the variance of $\hat{\tau}_p(h)$, but this happens at the cost of larger bias, and vice versa.

Theorem 1 provides an expression for the asymptotically optimal bandwidth h_{opt} that minimizes the AMSE of $\hat{\tau}_p(h)$. We can see that h_{opt} is increasing in the variation of the outcome at the cutoff and decreasing in the squared difference of the curvatures of the two conditional expectation functions at the cutoff. Furthermore, h_{opt} is decreasing in the sample size. The assumption $m_+^{(p+1)}(c) \neq m_-^{(p+1)}(c)$ when p is odd is made to avoid a case in which $B_p = 0$ and consequently $h_{opt} = \infty$. It is possible to derive the optimal bandwidth also for this setting by considering a higher order expansion of the bias of $\hat{\tau}_p(h)$. However, I leave this extension for future work.[9]

Theorem 6. *Suppose that $m_+^{(p+1)}(c) \neq m_-^{(p+1)}(c)$ when p is odd. Then*

$$\begin{aligned} MSE[\hat{\tau}_p(h)] &= AMSE[\hat{\tau}_p(h)] + o_p\left(h^{2(p+1)} + \frac{1}{nh}\right) \\ h_{opt} &= \arg\min_h AMSE[\hat{\tau}_p(h)] \\ &= C_{opt} n^{-\frac{1}{2p+3}} \end{aligned}$$

where

$$C_{opt} = \left(\frac{V_p}{2(p+1)B_p^2}\right)^{\frac{1}{2p+3}}.$$

Thus, the optimal bandwidth takes the form $h_{opt} = C_{opt} n^{-\frac{1}{2p+3}}$ for some constant $C_{opt} > 0$ that de-

[9] Arai and Ichimura (2013) propose a bandwidth choice algorithm for local linear regression-based estimator that takes into account the case $m_+^{(2)}(c) = m_-^{(2)}(c)$.

pends on the unknown parameters $m_+^{(p+1)}(c)$, $m_-^{(p+1)}(c)$, $\sigma_+^2(c)$, $\sigma_-^2(c)$, and $f(c)$. The problem of optimal bandwidth choice therefore boils down to the optimal choice of C in $h = Cn^{-\frac{1}{2p+3}}$. A common approach in the statistics and econometrics literature on LPR-based estimators is to estimate the unknown parameters that enter C_{opt}. In the RD design literature such plug-in estimator has been proposed for the local linear regression case by Imbens and Kalyanaraman (2012).[10]

3.3.2 Bandwidth Choice Algorithm

The bandwidth choice algorithm I propose in this paper is based on direct estimation of B_p^2 and V_p without the need to separately estimate the unknown parameters incorporated in these constants. The algorithm is general enough to be directly applicable to settings with arbitrary choices regarding the order of polynomial p and the kernel $K(u)$. The proposed approach also automatically adapts to various departures from the standard setting as discussed below.

The algorithm builds on the work by Schucany (1995) and Gerard and Schucany (1997).[11] The approach I take to estimate B_p^2 is motivated by the observation that

$$\begin{aligned}\hat{\tau}_p(h) - \tau &= B_p h^{p+1} + o_p\left(h^{p+1}\right) + O_p\left((nh)^{-\frac{1}{2}}\right) \\ \hat{\tau}_{p+1}(h) - \tau &= O\left(h^{p+2}\right) + o_p\left(h^{p+2}\right) + O_p\left((nh)^{-\frac{1}{2}}\right).\end{aligned}$$

That is, the leading term of the bias of $\hat{\tau}_{p+1}(h)$ is of higher order than that of $\hat{\tau}_p(h)$. Thus, letting $\hat{b}_p^2(h)$ denote the squared difference between these two estimators we get that

$$\hat{b}_p^2(h) = B_p^2 h^{2(p+1)} + o_p\left(h^{2(p+1)}\right) + O_p\left((nh)^{-1}\right).$$

The approach I take to estimate V_p is motivated by a similar observation as one can write the heteroskedastivity-robust variance estimator for $\hat{\tau}_p(h)$ as

$$\hat{v}_p(h) = \frac{V_p}{nh} + o_p\left(\frac{1}{nh}\right) + O_p\left((nh)^{-\frac{3}{2}}\right).$$

Taken together, these observations imply that one can estimate B_p^2 and V_p consistently by regressing $\hat{b}_p^2(h_k)$ on $h_k^{2(p+1)}$ and $\hat{v}_p(h)$ on $(nh_k)^{-1}$ using a collection of initial bandwidths h_k, $k = 1, \ldots, K$. The

[10] See also the alternative approaches to optimal bandwidth choice by Ludwig and Miller (2005) and DesJardins and McCall (2008) as well as the discussion regarding these approaches in Imbens and Kalyanaraman (2012).

[11] Similar approaches have also been proposed by Ruppert (1997), Doksum, Peterson, and Samarov (2000) and Prewitt (2003).

resulting Ordinary Least Squares (OLS) estimators of B_p^2 and V_p are

$$\begin{aligned}\hat{B}_p^2 &= \frac{\sum_{k=1}^{K} \hat{b}_p^2(h_k)\, h_k^{2(p+1)}}{\sum_{k=1}^{K} h_k^{4(p+1)}} \\ \hat{V}_p &= \frac{\sum_{k=1}^{K} \hat{v}_p(h_k)\,(nh_k)^{-1}}{\sum_{k=1}^{K} (nh_k)^{-2}}\end{aligned}$$

where the constant term in both regresions is restricted to zero.

By plugging in $\hat{B}_p^2$ and $\hat{V}_p$ to the expression for C_{opt} the estimator of the asymptotically optimal bandwidth h_{opt} becomes

$$\hat{h}_{opt} = \left(\frac{\hat{V}_p}{2(p+1)\hat{B}_p^2}\right)^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}.$$

The asymptotic properties of the bandwidth estimator $\hat{h}_{opt}$ and the resulting RD estimator $\hat{\tau}_p\left(\hat{h}_{opt}\right)$ are stated in Theorem 2. First, $\hat{h}_{opt}$ is a consistent estimator of the asymptotically optimal bandwidth h_{opt}. Second, the RD estimator $\hat{\tau}_p\left(\hat{h}_{opt}\right)$ satisfies the asymptotic optimality criterion of Li (1987). What this means is that, in terms of the MSE, the performance of the RD estimator using the estimated bandwidth is asymptotically as good as the performance of the RD estimator using the true optimal bandwidth. Third, the RD estimator $\hat{\tau}_p\left(\hat{h}_{opt}\right)$ converges to τ at the optimal nonparametric rate (Stone, 1982; Porter, 2003).

Theorem 7. *Suppose $h_k = c_k h^{-\gamma}$, $k = 1, \ldots, K$, for some positive, finite constants c_k and $\frac{1}{2p+5} \leq \gamma < \frac{1}{2p+3}$. Then*

$$\begin{aligned}\frac{\hat{h}_{opt}}{h_{opt}} - 1 &= o_p(1) \\ \frac{MSE\left[\hat{\tau}_p\left(\hat{h}_{opt}\right)\right]}{MSE\left[\hat{\tau}_p\left(h_{opt}\right)\right]} - 1 &= o_p(1) \\ \hat{\tau}_p\left(\hat{h}_{opt}\right) - \tau &= O_p\left(n^{-\frac{p+1}{2p+3}}\right).\end{aligned}$$

Note that the consistency of the bandwidth estimator $\hat{h}_{opt}$, and consequently the optimality properties of the resulting RD estimator $\hat{\tau}_p\left(\hat{h}_{opt}\right)$, require that the initial bandwidths used to estimate B_p^2 and V_p converge to zero at a slower rate than the asymptotically optimal bandwidth h_{opt}. While this is not necessary for the consistency of $\hat{V}_p$, it is needed to ensure the consistency of $\hat{B}_p$.

A remaining question is how one should choose the parameters c_k, γ, and K that define the collection of intial bandwidths h_k, $k = 1, \ldots, K$, in Theorem 1. Unfortunately, the asymptotic theory presented above has very little to say regarding these parameters. I propose to use the rate $\gamma = \frac{1}{2p+5}$ and the quantiles $0.50, 0.51, \ldots, 0.99$ of the distribution of $|R_i - c|$ as the constants c_k, $k = 1, \ldots, K$. It should be emphasized, however, that these choices are not motivated by any theoretical considerations. One could potentially improve the performance of the algorithm by using more appropriate parameter values, but I leave this

question for future research. Ideally, the resulting RD estimator $\hat{\tau}_p\left(\hat{h}_{opt}\right)$ is reasonably insensitive to these choices which is an important specification check when applying the algorithm.

3.4 Monte Carlo Experiments

In this section I compare the performance of the proposed bandwidth choice algorithm to the performance of the algorithm by Imbens and Kalyanaraman (2012). I follow Imbens and Kalyanaraman (2012) and explore finite sample behavior in Monte Carlo expriments that are based on the data from Lee (2008) who studies the effect of incumbency on the probability of re-election. As is common in empirical practice, I focus on the local linear regression-based estimator. Furthermore, I focus on the triangular kernel due to its optimality property mentioned above.

I consider the following functional forms for $m(r)$:[12]

$$
\begin{aligned}
m_1(r) &= \begin{cases} 0.48 + 1.27r + 7.18r^2 + 20.21r^3 + 21.54r^4 + 7.33r^5, r < 0 \\ 0.52 + 0.84r - 3.00r^2 + 7.99r^3 - 9.01r^4 + 3.56r^5, r \geq 0 \end{cases} \\
m_2(r) &= \begin{cases} 3r^2, r < 0 \\ 4r^2, r \geq 0 \end{cases} \\
m_3(r) &= \begin{cases} 0.42 + 0.84r - 3.00r^2 + 7.99r^3 - 9.01r^4 + 3.56r^5, r < 0 \\ 0.52 + 0.84r - 3.00r^2 + 7.99r^3 - 9.01r^4 + 3.56r^5, r \geq 0 \end{cases} \\
m_4(r) &= \begin{cases} 0.42 + 0.84r + 7.99r^3 - 9.01r^4 + 3.56r^5, r < 0 \\ 0.52 + 0.84r + 7.99r^3 - 9.01r^4 + 3.56r^5, r \geq 0 \end{cases}
\end{aligned}
$$

In all of the designs the running variable R_i and the residual ϵ_i are generated as $R_i \sim 2Beta(2,4) - 1$ and $\epsilon_i \sim N\left(0, 0.1295^2\right)$. I compare the behavior of the bandwidth choice algorithms in samples of size 100, 500, 1,000, 5,000, 10,000, and 50,000 using 1,000 replications.

The results from the Monte Carlo experiments are reported in Tables 3.1-3.4. The relative behavior of the two algorithms is similar across the different Monte Carlo designs and sample sizes. There are a few observations one can make based on these results. First, the adaptive algorithm tends to produce smaller bandwidths that vary somewhat more from one sample to another. Second, the bias of the RD estimator produced by the adative algorithm tends to be smaller. For the variance the situation is less clear: the adaptive algorithm tends to produce a less precise RD estimator in designs 1 and 3 while the opposite is true for designs 3 and 4. Finally, in terms of the MSE the adaptive algorithm performs better than the algorithm by Imbens and Kalyanaraman (2012) in designs 1, 2 and 4 once the sample size is at least 500 or 1,000 depending on the design. In design 2 the proposed algorithm performs instead worse than he algorithm by

[12]See the discussion in Imbens and Kalyanaraman (2012) regarding the choice of these functional forms.

Imbens and Kalyanaraman (2012) across all of the sample sizes.

Taken together, the results from the Monte Carlo experiments suggest that the adaptive algorithm has good finite-sample properties. This seems to be especially true for moderate sample sizes typically encountered in empirical applications. The proposed algorithm also compares well to the algorithm by Imbens and Kalyanaraman (2012).

3.5 Conclusions

This paper introduces an adaptive bandwidth choice algorithm for local polynomial regression-based estimators in the RD design. The algorithm is adaptive in the sense that it allows for different choices for the order of polynomial and kernel function. In addition, the algorithm automatically takes into account the inclusion of additional covariates as well as alternative assumptions on the variance-covariance structure of the error terms. I show that the algorithm produces a consistent estimator of the asymptotically optimal bandwidth that minimizes the AMSE as well as that the resulting RD estimator satisfies the asymptotic optimality criterion of Li (1987) and converges to the true parameter value at the optimal nonparametric convergence rate (Stone, 1982; Porter, 2003). Furthermore, Monte Carlo experiments suggest that the proposed algorithm has satisfactory finite-sample behavior and performs well in comparison to the algorithm by Imbens and Kalyanaraman (2012) for a local linear regression-based estimator.

I focus in the paper on sharp RD designs in which treatment is fully determined by the running variable. However, the proposed algorithm can be straightforwardly extended to fuzzy RD designs in which there is imperfect compliance with the treatment assignment. Another setting the approach can be applied to is the regression kink design (Card, Lee, Pei, and Weber, 2012) in which a continuous treatment variable has a kink instead of a discontinuity at a known cutoff. I leave these extensions for future research.

3.6 Tables

Table 3.1: Monte Carlo Simulations for Design 1

		$\hat{h}$		$\hat{\tau}$		
		Mean	SE	Bias	SE	RMSE
N = 100	IK	0.5637	0.1318	0.0335	0.0816	0.0882
	Adaptive	0.3302	0.0926	0.0294	0.1617	0.1643
N = 500	IK	0.4739	0.0585	0.0432	0.0359	0.0561
	Adaptive	0.2814	0.0751	0.0274	0.0525	0.0592
N = 1,000	IK	0.4161	0.0468	0.0423	0.0245	0.0489
	Adaptive	0.2477	0.0704	0.0217	0.0397	0.0453
N = 5,000	IK	0.3399	0.0337	0.0385	0.0110	0.0400
	Adaptive	0.1671	0.0443	0.0136	0.0210	0.0250
N = 10,000	IK	0.3311	0.0266	0.0380	0.0086	0.0390
	Adaptive	0.1379	0.0337	0.0107	0.0166	0.0197
N = 50,000	IK	0.1988	0.0184	0.0223	0.0070	0.0234
	Adaptive	0.0883	0.0118	0.0054	0.0083	0.0099

Table 3.2: Monte Carlo Simulations for Design 2

		$\hat{h}$		$\hat{\tau}$		
		Mean	SE	Bias	SE	RMSE
N = 100	IK	0.5581	0.1535	0.0287	0.0926	0.0969
	Adaptive	0.3317	0.0970	0.0045	0.1635	0.1636
N = 500	IK	0.4189	0.0712	0.0087	0.0369	0.0379
	Adaptive	0.3012	0.0802	0.0015	0.0506	0.0506
N = 1,000	IK	0.3643	0.0472	0.0025	0.0259	0.0261
	Adaptive	0.2699	0.0701	0.0004	0.0371	0.0371
N = 5,000	IK	0.2624	0.0203	0.0016	0.0137	0.0138
	Adaptive	0.2191	0.0566	0.0012	0.0176	0.0176
N = 10,000	IK	0.2285	0.0167	0.0017	0.0109	0.0111
	Adaptive	0.1980	0.0507	0.0009	0.0135	0.0135
N = 50,000	IK	0.1661	0.0089	0.0014	0.0054	0.0055
	Adaptive	0.1572	0.0410	0.0008	0.0063	0.0064

Table 3.3: Monte Carlo Simulations for Design 3

		$\hat{h}$		$\hat{\tau}$		
		Mean	SE	Bias	SE	RMSE
N = 100	IK	0.2289	0.0254	0.0234	0.1289	0.1310
	Adaptive	0.2105	0.0396	0.0169	0.1770	0.1778
N = 500	IK	0.1746	0.0160	0.0078	0.0582	0.0587
	Adaptive	0.1802	0.0263	0.0067	0.0581	0.0585
N = 1,000	IK	0.1563	0.0143	0.0057	0.0442	0.0445
	Adaptive	0.1643	0.0229	0.0049	0.0439	0.0442
N = 5,000	IK	0.1226	0.0107	0.0030	0.0210	0.0212
	Adaptive	0.1327	0.0182	0.0031	0.0209	0.0211
N = 10,000	IK	0.1106	0.0099	0.0021	0.0160	0.0162
	Adaptive	0.1209	0.0163	0.0022	0.0158	0.0159
N = 50,000	IK	0.0877	0.0077	0.0011	0.0076	0.0077
	Adaptive	0.0970	0.0134	0.0012	0.0075	0.0076

Table 3.4: Monte Carlo Simulations for Design 4

		$\hat{h}$		$\hat{\tau}$		
		Mean	SE	Bias	SE	RMSE
N = 100	IK	0.2238	0.0251	0.0192	0.1315	0.1329
	Adaptive	0.2157	0.0399	0.0155	0.1774	0.1781
N = 500	IK	0.1735	0.0165	0.0063	0.0581	0.0584
	Adaptive	0.1873	0.0283	0.0060	0.0573	0.0576
N = 1,000	IK	0.1556	0.0144	0.0047	0.0441	0.0444
	Adaptive	0.1710	0.0250	0.0044	0.0432	0.0435
N = 5,000	IK	0.1225	0.0108	0.0026	0.0210	0.0211
	Adaptive	0.1386	0.0208	0.0029	0.0205	0.0207
N = 10,000	IK	0.1105	0.0100	0.0018	0.0160	0.0161
	Adaptive	0.1262	0.0186	0.0021	0.0156	0.0157
N = 50,000	IK	0.0877	0.0077	0.0010	0.0076	0.0077
	Adaptive	0.1015	0.0154	0.0011	0.0074	0.0075

3.7 Appendix

Preliminaries

The estimator of $m_+(c)$ using a p^{th} order polynomial and a bandwidth h,$\hat{\alpha}_p^+(h)$, is obtained by solving

$$\min_{\alpha,\{\beta_k\}_{k=1}^p} \sum_{i=1}^N K_+\left(\frac{R_i-c}{h}\right)\left(Y_i-\alpha-\sum_{k=1}^p \beta_k (R_i-c)^k\right)^2$$

where $K_+(u) = 1(u \geq 0) K(u)$. Let $\alpha^+ = m_+(c)$ and $\beta_k^+ = m_+^{(k)}(c)$, $k = 1,\ldots,p$, and define $U_i = Y_i - \alpha^+ - \sum_{k=1}^p \beta_k^+ (R_i-c)^k$. Using this notation the minimization problem can be rewritten as

$$\min_{(\alpha-\alpha^+),\left\{h^k\left(\beta_k-\beta_k^+\right)\right\}_{k=1}^p} \sum_{i=1}^N 1(R_i \geq c) K_+\left(\frac{R_i-c}{h}\right)\left(U_i-(\alpha-\alpha^+)-\sum_{k=1}^p h^k\left(\beta_k-\beta_k^+\right)\left(\frac{R_i-c}{h}\right)^k\right)^2.$$

From the first order conditions we get

$$\begin{bmatrix} \hat{\alpha}_p^+(h)-\alpha^+ \\ h\left(\hat{\beta}_{1,p}^+(h)-\beta_1^+\right) \\ \vdots \\ h^p\left(\hat{\beta}_{p,p}^+(h)-\beta_p^+\right) \end{bmatrix} = \left[\frac{1}{nh_n}\sum_{i=1}^N K_+\left(\frac{R_i-c}{h}\right) X_{pi}X_{pi}'\right]^{-1}\left[\frac{1}{nh}\sum_{i=1}^N K_+\left(\frac{R_i-c}{h}\right) X_{pi}U_i\right]$$

where $X_{pi} = \begin{bmatrix} 1 & \left(\frac{R_i-c}{h}\right) & \cdots & \left(\frac{R_i-c}{h}\right)^p \end{bmatrix}'$.

Finally, define

$$\xi(r) = m(r) - \alpha^+ - \sum_{k=1}^p \beta_k^+ (r-c)^p - \frac{m_+^{(p+1)}(c)}{(p+1)!}(r-c)^{p+1}$$

and note that

$$\sup_{r\in(c,c+Mh]} |\xi(r)| = O\left(h^{p+2}\right).$$

Lemma 1

$$\frac{1}{nh}\sum_{i=1}^N K_+\left(\frac{R_i-c}{h}\right) X_{pi}X_{pi}' = f(c)\Gamma_+ + o(1) + o_p(1)$$

where

$$\Gamma_+ = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_p \\ \gamma_1 & \gamma_2 & \cdots & \gamma_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_p & \gamma_{p+1} & \cdots & \gamma_{2p} \end{bmatrix}$$

and

$$\gamma_k = \int_0^\infty u^k K(u)\, du.$$

Proof: Let us write

$$\frac{1}{nh}\sum_{i=1}^{N} K_+\left(\frac{R_i - c}{h}\right) X_{pi} X_{pi}' = \begin{bmatrix} A_{0,n} & A_{1,n} & \cdots & A_{p,n} \\ A_{1,n} & A_{2,n} & \cdots & A_{p+1,n} \\ \vdots & \vdots & \ddots & \vdots \\ A_{p,n} & A_{p+1,n} & \cdots & A_{2p,n} \end{bmatrix}$$

where

$$A_{k,n} = \frac{1}{nh}\sum_{i=1}^{N} K_+\left(\frac{R_i - c}{h}\right)\left(\frac{R_i - c}{h}\right)^k.$$

For the mean of $A_{k,n}$ we get

$$\begin{aligned} E[A_{k,n}] &= \frac{1}{h} E\left[K_+\left(\frac{R_i - c}{h}\right)\left(\frac{R_i - c}{h}\right)^k\right] \\ &= \frac{1}{h}\int_c^\infty K\left(\frac{r-c}{h}\right)\left(\frac{r-c}{h}\right)^k f(r)\, dr \\ &= \int_0^\infty K(u)\, u^k f(c + hu)\, du \\ &= f(c)\int_0^\infty u^k K(u)\, du + o(1) \end{aligned}$$

where the third equality follows from a change of variables $u = \frac{r-c}{h}$ and the fourth equality from the dominated convergence theorem.

For the variance of $A_{k,n}$ we get

$$\begin{aligned} Var\left[A_{k,n}\right] &\leq \frac{1}{nh^2} E\left[K_+\left(\frac{R_i-c}{h}\right)^2\left(\frac{R_i-c}{h}\right)^{2k}\right] \\ &= \frac{1}{nh^2}\int_c^\infty K\left(\frac{r-c}{h}\right)^2\left(\frac{r-c}{h}\right)^{2k} f(r)\,dr \\ &= \frac{1}{nh}\int_0^\infty K(u)^2 u^{2k} f(c+hu)\,du \\ &= o(1) \end{aligned}$$

where the second equality follows from a change of variables $u = \frac{r-c}{h}$ and the third equality from the dominated convergence theorem. □

Lemma 2

$$E\left[\frac{1}{nh}\sum_{i=1}^N K_+\left(\frac{R_i-c}{h}\right)X_{pi}U_i\right] = \frac{m_+^{(p+1)}(c)}{(p+1)!} f(c)\,\delta_+ h^{p+1} + O\left(h^{p+2}\right)$$

where

$$\delta_+ = \begin{bmatrix} \delta_0 \\ \vdots \\ \delta_p \end{bmatrix}$$

and

$$\delta_k = \int_0^\infty u^{k+p+1} K(u)\,du$$

Proof: Let

$$\frac{1}{nh}\sum_{i=1}^N K_+\left(\frac{R_i-c}{h}\right)X_{pi}U_i = \begin{bmatrix} A_{0,n} \\ \vdots \\ A_{p,n} \end{bmatrix}$$

where

$$\begin{aligned} A_{k,n} &= \frac{1}{nh}\sum_{i=1}^{n} K_+\left(\frac{R_i-c}{h}\right)\left(\frac{R_i-c}{h}\right)^k U_i \\ &= \frac{1}{nh}\sum_{i=1}^{n} K_+\left(\frac{R_i-c}{h}\right)\left(\frac{R_i-c}{h}\right)^k \left(\frac{m_+^{(p+1)}(c)}{(p+1)!}(R_i-c)^{p+1}+\xi(R_i)+\epsilon_i\right). \end{aligned}$$

We can now write

$$\begin{aligned} E[A_{k,n}] &= \frac{1}{h}E\left[K_+\left(\frac{R_i-c}{h}\right)\left(\frac{R_i-c}{h}\right)^k\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}(R_i-c)^{p+1}+\xi(R_i)\right)\right] \\ &= \frac{m_+^{(p+1)}(c)}{(p+1)!}h^p E\left[K_+\left(\frac{R_i-c}{h}\right)\left(\frac{R_i-c}{h}\right)^{k+p+1}\right] \\ &\quad +\frac{1}{h}E\left[K_+\left(\frac{R_i-c}{h}\right)\left(\frac{R_i-c}{h_n}\right)^k \xi(R_i)\right] \\ &= \frac{m_+^{(p+1)}(c)}{(p+1)!}h^p\int_c^\infty K\left(\frac{r-c}{h}\right)\left(\frac{r-c}{h}\right)^{k+p+1} f(r)\,dr \\ &\quad +O\left(h^{p+1+\eta}\right)\frac{1}{h_n}\int_c^\infty K\left(\frac{r-c}{h_n}\right)\left(\frac{r-c}{h_n}\right)^k f(r)\,dr \\ &= \frac{m_+^{(p+1)}(c)}{(p+1)!}h^{p+1}\int_0^\infty K(u)\,u^{k+p+1} f(c+hu)\,du \\ &\quad +O\left(h^{p+2}\right)\int_0^\infty K(u)\,u^k f(c+hu)\,du \\ &= \frac{m_+^{(p+1)}(c)}{(p+1)!}f(c)\,h^{p+1}\left(\int_0^\infty u^{k+p+1}K(u)\,du+o(1)\right)+O\left(h^{p+2}\right). \end{aligned}$$

where the fourth equality from a change of variables $u=\frac{r-c}{h_n}$ and the fifth equality from the dominated convergence theorem. □

Lemma 3

$$Var\left[\frac{1}{nh}\sum_{i=1}^{n} K_+\left(\frac{R_i-c}{h}\right)X_{pi}U_i\right] = \frac{1}{nh}\left(\sigma_+^2(c)\,f(c)\,\Lambda_+ + o(1)+O\left(h^{2(p+1)}\right)\right)$$

where

$$\Lambda_+ = \begin{bmatrix} \lambda_0 & \lambda_1 & \cdots & \lambda_p \\ \lambda_1 & \lambda_2 & \cdots & \lambda_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_p & \lambda_{p+1} & \cdots & \lambda_{2p} \end{bmatrix}$$

and

$$\lambda_k = \int_0^\infty u^k K(u)^2 \, du.$$

Proof: Note that

$$\begin{aligned} Var\left[\frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right) X_{pi} U_i\right] &= E\left[Var\left[\frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right) X_{pi} U_i \mid X_{pi}\right]\right] \\ &\quad + Var\left[E\left[\frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right) X_{pi} U_i \mid X_{pi}\right]\right]. \end{aligned}$$

Let us first look at

$$E\left[Var\left[\frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right) X_{pi} U_i \mid X_{pi}\right]\right] = E\left[Var\left[\frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right) X_{pi} \epsilon_i \mid X_{pi}\right]\right].$$

I will only consider the variance of

$$A_{k,n} = \frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right)\left(\frac{R_i - c}{h}\right)^k \epsilon_i$$

as the covariance terms can be handled in a similar fashion. For this we get

$$\begin{aligned} E\left[Var\left[A_{k,n} \mid X_{pi}\right]\right] &= \frac{1}{nh^2} E\left[K_+\left(\frac{R_i - c}{h}\right)^2 \left(\frac{R_i - c}{h}\right)^{2k} \sigma^2(R_i)\right] \\ &= \frac{1}{nh^2}\int_c^\infty K\left(\frac{r - c}{h}\right)^2 \left(\frac{r - c}{h}\right)^{2k} \sigma^2(r) f(r) \, dr \\ &= \frac{1}{nh}\int_0^\infty K(u)^2 u^{2k} \sigma^2(c + hu) f(c + hu) \, du \\ &= \frac{1}{nh}\sigma_+^2(c) f(c)\left(\int_0^\infty u^{2k} K(u)^2 \, du + o(1)\right) \end{aligned}$$

where the fourth equality follows from a change of variables $u = \frac{r-c}{h}$ and the fifth equality from the dominated convergence theorem.

Let us now turn to the second term and note that

$$E\left[\frac{1}{nh}\sum_{i=1}^{n} K_+\left(\frac{R_i-c}{h}\right) X_{pi}U_i \mid X_{pi}\right] = \frac{1}{nh}\sum_{i=1}^{n} K_+\left(\frac{R_i-c}{h}\right) X_{pi}\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}(R_i-c)^{p+1} + \xi(R_i)\right).$$

I will only consider the variance of

$$A_{k,n} = \frac{1}{nh}\sum_{i=1}^{n} K_+\left(\frac{R_i-c}{h}\right)\left(\frac{R_i-c}{h}\right)^k\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}(R_i-c)^{p+1} + \xi(R_i)\right)$$

as the covariance terms can be handled in a similar fashion. For this we get

$$\begin{aligned}
Var[A_{k,n}] &\leq \frac{1}{nh^2}E\left[K_+\left(\frac{R_i-c}{h}\right)^2\left(\frac{R_i-c}{h}\right)^{2k}\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}(R_i-c)^{p+1} + \xi(R_i)\right)^2\right] \\
&\leq \frac{O(1)}{nh^2}E\left[K_+\left(\frac{R_i-c}{h}\right)^2\left(\frac{R_i-c}{h}\right)^{2k}\left(\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}\right)^2(R_i-c)^{2(p+1)} + \xi(R_i)^2\right)\right] \\
&= \frac{O(1)}{nh^2}h^{2(p+1)}\int_c^{\infty} K\left(\frac{r-c}{h}\right)^2\left(\frac{r-c}{h}\right)^{2(k+p+1)} f(r)\,dr \\
&\quad + \frac{O(1)}{nh^2}O\left(h^{2(p+2)}\right)\int_c^{\infty} K\left(\frac{r-c}{h}\right)^2\left(\frac{r-c}{h}\right)^{2k} f(r)\,dr \\
&= \frac{O(1)}{nh}h^{2(p+1)}\int_0^{\infty} K(u)^2 u^{2(k+p+1)} f(c+hu)\,du \\
&\quad + \frac{O(1)}{nh}O\left(h^{2(p+2)}\right)\int_0^{\infty} K(u)^2 u^{2k} f(c+hu)\,du \\
&= \frac{1}{nh}O\left(h^{2(p+1)}\right)
\end{aligned}$$

where the fourth equality follows from a change of variables $u = \frac{r-c}{h}$ and the fifth equality from the dominated convergence theorem. □

Lemma 4

$$E\begin{bmatrix} \hat{\alpha}_p^+(h) - \alpha^+ \\ h\left(\hat{\beta}_{1,p}(h) - \beta_1^+\right) \\ \vdots \\ h^p\left(\hat{\beta}_{p,p}(h) - \beta_p^+\right) \end{bmatrix} = \frac{m_+^{(p+1)}(c)}{(p+1)!}\Gamma_+^{-1}\delta_+ h^{p+1} + o_p\left(h^{p+1}\right)$$

$$Var\begin{bmatrix} \hat{\alpha}_p^+(h) - \alpha^+ \\ h\left(\hat{\beta}_{1,p}(h) - \beta_1^+\right) \\ \vdots \\ h^p\left(\hat{\beta}_{p,p}(h) - \beta_p^+\right) \end{bmatrix} = \frac{1}{nh}\frac{\sigma_+^2(c)}{f(c)}\Gamma_+^{-1}\Lambda_+\Gamma_+^{-1} + o_p\left(\frac{1}{nh}\right)$$

Proof: The result follows from Lemma 1, Lemma 2 and Lemma 3 using the continuous mapping theorem. □

Lemma 5

$$\frac{1}{nh_n}\sum_{i=1}^{n} K_+\left(\frac{R_i - c}{h_n}\right)^2 U_i^2 X_{pi} X_{pi}' = \sigma_+^2(c) f(c) \Lambda_+ + O\left(h^{2(p+1)}\right) + o_p(1)$$

where

$$\Lambda_+ = \begin{bmatrix} \lambda_0 & \lambda_1 & \cdots & \lambda_p \\ \lambda_1 & \lambda_2 & \cdots & \lambda_{p+1} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_p & \lambda_{p+1} & \cdots & \lambda_{2p} \end{bmatrix}$$

and

$$\lambda_k = \int_0^\infty u^k K(u)^2 du.$$

Proof: Let us start by writing

$$\frac{1}{nh_n}\sum_{i=1}^{n} K_+\left(\frac{R_i - c}{h_n}\right)^2 U_i^2 X_{pi} X_{pi}' = \begin{bmatrix} A_{0,n} & \cdots & A_{p,n} \\ \vdots & \ddots & \vdots \\ A_{p,n} & \cdots & A_{2p,n} \end{bmatrix}$$

where

$$\begin{aligned} A_{k,n} &= \frac{1}{nh_n}\sum_{i=1}^{n} K_+\left(\frac{R_i-c}{h_n}\right)^2\left(\frac{R_i-c}{h_n}\right)^k U_i^2 \\ &= \frac{1}{nh}\sum_{i=1}^{n} K_+\left(\frac{R_i-c}{h}\right)^2\left(\frac{R_i-c}{h}\right)^k\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}(R_i-c)^{p+1}+\xi(R_i)+\epsilon_i\right)^2. \end{aligned}$$

The expectation of $A_{k,n}$ can be written as

$$\begin{aligned} E[A_{k,n}] &= \frac{1}{h}E\left[K_+\left(\frac{R_i-c}{h}\right)^2\left(\frac{R_i-c}{h}\right)^k\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}(R_i-c)^{p+1}+\xi(R_i)+\epsilon_i\right)^2\right] \\ &= \frac{1}{h}E\left[K_+\left(\frac{R_i-c}{h_n}\right)^2\left(\frac{R_i-c}{h_n}\right)^k\sigma^2(R_i)\right]+R_n \\ &= \frac{1}{h}\int_c^\infty K\left(\frac{r-c}{h}\right)^2\left(\frac{r-c}{h}\right)^k\sigma^2(r)f(r)\,dr+R_n \\ &= \int_0^\infty K(u)^2u^k\sigma^2(c+hu)f(c+hu)\,du+R_n \\ &= \sigma_+^2(c)f(c)\int_0^\infty u^kK(u)^2\,du+o(1)+R_n \end{aligned}$$

where the fourth equality from a change of variables $u=\frac{r-c}{h}$ and the fifth equality from the dominated convergence theorem. Furthermore, notice that

$$\begin{aligned} R_n &= \frac{1}{h}E\left[K_+\left(\frac{R_i-c}{h}\right)^2\left(\frac{R_i-c}{h_n}\right)^k\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}(R_i-c)^{p+1}+\xi(R_i)\right)^2\right] \\ &\le O(1)\frac{1}{h}E\left[K_+\left(\frac{R_i-c}{h}\right)^2\left(\frac{R_i-c}{h}\right)^k\left(\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}\right)^2(R_i-c)^{2(p+1)}+\xi(R_i)^2\right)\right] \\ &= O(1)\frac{1}{h}\int_c^\infty K\left(\frac{r-c}{h}\right)^2\left(\frac{r-c}{h}\right)^k(R_i-c)^{2(p+1)}f(r)\,dr \\ &\quad +O(1)\frac{1}{h}\int_0^\infty K\left(\frac{r-c}{h}\right)^2\left(\frac{r-c}{h}\right)^k\xi(R_i)^2f(r)\,dr \\ &= O(1)h^{2(p+1)}\int_0^\infty K(u)^2u^{k+2(p+1)}f(c+hu)\,du \\ &\quad +O(1)O\left(h^{2(p+2)}\right)\int_0^\infty K(u)^2u^kf(c+hu)\,du \\ &= O\left(h^{2(p+1)}\right) \end{aligned}$$

where the first inequality follows from the c_r inequality and the third equality from a change of variables $u = \frac{r-c}{h}$.

For the variance of $A_{k,n}$ we get

$$\begin{aligned}
Var\left[A_{k,n}\right] &\leq \frac{1}{nh^2}E\left[K_+\left(\frac{R_i-c}{h}\right)^4\left(\frac{R_i-c}{h}\right)^{2k}\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}(R_i-c)^{p+1}+\xi(R_i)+\epsilon_i\right)^4\right] \\
&\leq \frac{1}{nh^2}O(1)E\left[K_+\left(\frac{R_i-c}{h}\right)^4\left(\frac{R_i-c}{h}\right)^{2k}\left(\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}\right)^4(R_i-c)^{4(p+1)}+\xi(R_i)^4+\epsilon_i^4\right)\right] \\
&= \frac{1}{nh^2}O(1)\int_c^\infty K\left(\frac{r-c}{h}\right)^4\left(\frac{r-c}{h}\right)^{2k}\left(\frac{m_+^{(p+1)}(c)}{(p+1)!}\right)^4(r-c)^{4(p+1)}f(r)\,dr \\
&\quad +\frac{1}{nh^2}O(1)\int_c^\infty K\left(\frac{r-c}{h}\right)^4\left(\frac{r-c}{h}\right)^{2k}\xi(r)^4 f(r)\,dr \\
&\quad +\frac{1}{nh^2}O(1)\int_c^\infty K\left(\frac{r-c}{h}\right)^4\left(\frac{r-c}{h}\right)^{2k}E\left[\epsilon_i^4 \mid R_i=r\right]f(r)\,dr \\
&= \frac{1}{nh}O(1)h^{4(p+1)}\int_0^\infty K(u)^4 u^{2k+4(p+1)}f(c+hu)\,du \\
&\quad +\frac{1}{nh}O(1)o\left(h^{4(p+1)}\right)\int_0^\infty K(u)^4(u)^{2k}f(c+hu)\,du \\
&\quad +\frac{1}{nh}O(1)\int_0^\infty K(u)^4 u^{2k}E\left[\epsilon_i^4 \mid R_i=c+hu\right]f(c+hu)\,du \\
&= O\left(\frac{1}{nh}\right)
\end{aligned}$$

where the second inequality follows from the c_r inequality and the second equality from a change of variables $u = \frac{r-c}{h}$. □

Lemma 6

$$\frac{1}{nh_n}\sum_{i=1}^n K_+\left(\frac{R_i-c}{h_n}\right)^2\hat{U}_i^2 X_{pi}X_{pi}' \quad = \frac{1}{nh_n}\sum_{i=1}^n K_+\left(\frac{R_i-c}{h_n}\right)^2 U_i^2 X_{pi}X_{pi}' + \quad o_p(1)$$

where

$$\hat{U}_i \quad = \quad Y_i - \hat{\alpha}_p^+(h) - \sum_{k=1}^p \hat{\beta}_{p,k}^+(h)(R_i-c)^k$$

Proof: Let us start by writing

$$\begin{aligned}\hat{U}_i^2 &= \left(U_i + \hat{U}_i - U_i\right)^2 \\ &= U_i^2 + \left(\hat{U}_i - U_i\right)^2 + 2U_i\left(\hat{U}_i - U_i\right) \\ &= U_i^2 + \left(\tilde{\beta}_p^+(h)' X_{pi}\right)^2 + 2U_i\left(\tilde{\beta}_p^+(h)' X_{pi}\right)\end{aligned}$$

where $\tilde{\beta}_p^+(h) = \left[\hat{\alpha}_p^+(h) - \alpha^+, h\left(\hat{\beta}_{1,p}(h) - \beta_1^+\right), \ldots, h^p\left(\hat{\beta}_{p,p}(h) - \beta_p^+\right)\right]'$. Note that $\tilde{\beta}_p^+(h) = o_p(1)$ by Lemma 4. We can now write

$$\frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right)^2 \hat{U}_i^2 X_{pi} X_{pi}' = \begin{bmatrix} A_{0,n} & \cdots & A_{p,n} \\ \vdots & \ddots & \vdots \\ A_{p,n} & \cdots & A_{2p,n}\end{bmatrix}$$

where

$$\begin{aligned}A_{k,n} &= \frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right)^2\left(\frac{R_i - c}{h}\right)^k \hat{U}_i^2 \\ &= \frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right)^2\left(\frac{R_i - c}{h}\right)^k U_i^2 + R_n\end{aligned}$$

and

$$\begin{aligned}R_n &= \sum_{l=0}^p \frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right)^2\left(\frac{R_i - c}{h}\right)^k\left(\tilde{\beta}_p^+(h)' X_{pi}\right)^2 \\ &\quad + O(1)\sum_{l=0}^p \frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right)^2\left(\frac{R_i - c}{h}\right)^{k+2l} U_i\left(\tilde{\beta}_p^+(h)' X_{pi}\right) \\ &\leq o_p(1)\frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right)^2\sum_{l=0}^p\left(\frac{R_i - c}{h}\right)^{k+2l} \\ &\quad + o_p(1)\frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right)^2\sum_{l=0}^p\left(\frac{R_i - c}{h}\right)^{k+l} U_i\end{aligned}$$

where the inequality follows from the c_r inequality. The result follows from observing that

$$\begin{aligned}\frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right)^2\sum_{l=0}^p\left(\frac{R_i - c}{h}\right)^{k+2l} &= O(1) + o_p(1) \\ \frac{1}{nh}\sum_{i=1}^n K_+\left(\frac{R_i - c}{h}\right)^2\sum_{l=0}^p\left(\frac{R_i - c}{h}\right)^{k+l} U_i &= O(1) + o_p(1)\end{aligned}$$

which can be shown using similar steps as above. □

Lemma 7

$$E\left[v_p^+(h)\right] = \frac{1}{nh}\frac{\sigma_+^2(c)}{f(c)}\Gamma_+^{-1}\Lambda_+\Gamma_+^{-1} + o_p\left(\frac{1}{nh}\right)$$

Proof: The result follows from Lemma 1, Lemma 5 and Lemma 6 using the continuous mapping theorem. □

Proof of Theorem 1

From Lemma 4 and similar result for the LPR estimator using observations to the left of the cutoff we get that

$$\begin{aligned} Bias\left[\hat{\tau}_p(h)\right] &= B_p h^{p+1} + o_p\left(h^{p+1}\right) \\ Var\left[\hat{\tau}_p(h)\right] &= \frac{V_p}{nh} + o_p\left(\frac{1}{nh}\right) \end{aligned}$$

where

$$\begin{aligned} B_p &= \frac{m_+^{(p+1)}(c)}{(p+1)!}\Gamma_+^{-1}\delta_+ - \frac{m_-^{(p+1)}(c)}{(p+1)!}\Gamma_-^{-1}\delta_- \\ V_p &= \frac{\sigma_+^2(c)}{f(c)}\Gamma_+^{-1}\Lambda_+\Gamma_+^{-1} - \frac{\sigma_-^2(c)}{f(c)}\Gamma_-^{-1}\Lambda_-\Gamma_-^{-1}. \end{aligned}$$

Thus, we have that

$$\begin{aligned} MSE\left[\hat{\tau}_p(h)\right] &= Bias\left[\hat{\tau}_p(h)\right]^2 + Var\left[\hat{\tau}_p(h)\right] \\ &= AMSE\left[\hat{\tau}_p(h)\right] + o_p\left(h^{2(p+1)} + \frac{1}{nh}\right) \end{aligned}$$

where

$$AMSE\left[\hat{\tau}_p(h)\right] = B_p^2 h^{2(p+1)} + \frac{V_p}{nh}.$$

Finally, note that $AMSE\left[\hat{\tau}_p(h)\right]$ is globally convex in h. Thus differentiating this with respect to h and taking the first order condition gives us

$$\begin{aligned} h_{opt} &= \arg\min_h AMSE\left[\hat{\tau}_p(h)\right] \\ &= \left(\frac{V_p}{2(p+1)B_p^2}\right)^{\frac{1}{2p+3}} n^{-\frac{1}{2p+3}}. \end{aligned}$$

□

Proof of Theorem 2

From Lemma 4 and similar result for the LPR estimator using observations to the left of the cutoff as well corresponding results for a LPR-based estimator using a $(p+1)^{th}$ order polynomial we get that

$$\begin{aligned}\hat{\tau}_p(h) &= \tau + B_p h^{p+1} + o_p\left(h^{p+1}\right) + O_p\left((nh)^{-\frac{1}{2}}\right) \\ \hat{\tau}_{p+1}(h) &= \tau + o_p\left(h^{p+1}\right) + O_p\left((nh)^{-\frac{1}{2}}\right).\end{aligned}$$

This implies that

$$\hat{\tau}_p(h) - \hat{\tau}_{p+1}(h) = B_p h^{p+1} + o_p\left(h^{p+1}\right) + O_p\left((nh)^{-\frac{1}{2}}\right)$$

and consequently that

$$\begin{aligned}\hat{b}_p^2(h) &= \left(\hat{\tau}_p(h) - \hat{\tau}_{p+1}(h)\right) \\ &= B_p^2 h^{2(p+1)} + o_p\left(h^{2(p+1)}\right) + O_p\left((nh)^{-1}\right).\end{aligned}$$

In addition, from Lemma 6 and similar result for the variance estimator to the left of the cutoff we get that

$$\hat{v}_p(h) = \frac{V_p}{nh} + o_p\left(\frac{1}{nh}\right) + O_p\left((nh)^{-\frac{3}{2}}\right).$$

Thus, by plugging in h_k we get

$$\begin{aligned}\hat{b}_p^2(h_k) h_k^{-2(p+1)} &= B_p^2 + o_p(1) + O_p\left(n^{-(1-2p+3)\gamma}\right) \\ \hat{v}_p(h_k) nh_k &= V_p + o_p(1) + O_p\left(n^{-\frac{1}{2}(1-\gamma)}\right).\end{aligned}$$

This implies that

$$\begin{aligned}\hat{B}_p^2 &= \frac{\sum_{k=1}^K \hat{b}_p^2(h_k) h_k^{2(p+1)}}{\sum_{k=1}^K h_k^{4(p+1)}} \\ &= \frac{\sum_{k=1}^K \hat{b}_p^2(h_k) h_k^{-2(p+1)} h_k^{4(p+1)}}{\sum_{k=1}^K h_k^{4(p+1)}} \\ &= B_p^2 + o_p(1) + O_p\left(n^{-(1-2p+3)\gamma}\right)\end{aligned}$$

and that

$$\begin{aligned}\hat{V}_p &= \frac{\sum_{k=1}^{K}\hat{v}_p(h_k)(nh_k)^{-1}}{\sum_{k=1}^{K}(nh_k)^{-2}} \\ &= \frac{\sum_{k=1}^{K}\hat{v}_p(h_k)nh_k(nh_k)^{-2}}{\sum_{k=1}^{K}(nh_k)^{-2}} \\ &= V_p + o_p(1) + O_p\left(n^{-\frac{1}{2}(1-\gamma)}\right).\end{aligned}$$

Furthermore, plugging in the estimators $\hat{B}_p$ and $\hat{V}_p$ to the expression for C_{opt}, we get

$$\begin{aligned}\hat{C}_{opt} &= \left(\frac{\hat{V}_p}{2(p+1)\hat{B}_p^2}\right)^{\frac{1}{2p+3}} \\ &= \left(\frac{V_p}{2(p+1)B_p^2}\right)^{\frac{1}{2p+3}} + o_p(1) \\ &= C_{opt} + o_p(1)\end{aligned}$$

and consequently that

$$\begin{aligned}\frac{\hat{h}_{opt}}{h_{opt}} &= \frac{\hat{C}_{opt}n^{-\frac{1}{2p+3}}}{C_{opt}n^{-\frac{1}{2p+3}}} \\ &= \frac{\hat{C}_{opt}}{C_{opt}} \\ &= 1 + o_p(1).\end{aligned}$$

Finally, note that

$$\begin{aligned}MSE\left[\hat{\tau}_p(h_{opt})\right] &= AMSE\left[\hat{\tau}_p(h_{opt})\right] + o_p\left(h_{opt}^{2(p+1)} + \frac{1}{nh_{opt}}\right) \\ &= n^{-\frac{2(p+1)}{2[p+3}}\left(B_pC_{opt}^{2(p+1)} + V_pC_{opt}^{-1} + o_p(1)\right)\end{aligned}$$

and that

$$\begin{aligned}MSE\left[\hat{\tau}_p\left(\hat{h}_{opt}\right)\right] &= AMSE\left[\hat{\tau}_p\left(\hat{h}_{opt}\right)\right] + o_p\left(\hat{h}_{opt}^{2(p+1)} + \frac{1}{n\hat{h}_{opt}}\right) \\ &= n^{-\frac{2(p+1)}{2[p+3}}\left(B_pC_{opt}^{2(p+1)} + V_pC_{opt}^{-1} + o_p(1)\right).\end{aligned}$$

Thus, we get that

$$\frac{MSE\left[\hat{\tau}_p\left(\hat{h}_{opt}\right)\right]}{MSE\left[\hat{\tau}_p(h_{opt})\right]} = 1 + o_p(1)$$

which implies that

$$\hat{\tau}_p\left(\hat{h}_{opt}\right) \quad = \quad \tau + O_p\left(n^{-\frac{p+1}{2(p+3}}\right).$$

□

Bibliography

AAKVIK, A., J. HECKMAN, AND E. VYTLACIL (2005): "Estimating Treatment Effects for Discrete Outcomes When Responses to Treatment Vary: An Application to Norwegian Vocational Rehabilitation Programs," *Journal of Econometrics*, 125(1-2), 15–51.

ABADIE, A. (2003): "Semiparametric Instrumental Variables Estimation of Treatment Response Models," *Journal of Econometrics*, 113(2), 231–263.

——— (2005): "Semiparametric Difference-in-Difference Estimators," *Review of Economic Studies*, 72(1), 1–19.

ABADIE, A., J. ANGRIST, AND G. IMBENS (2002): "Instrumental Variables Estimates of the Effect of Subsidized Training on the Quantiles of Trainee Earnings," *Econometrica*, 70(1), 91–117.

ABDULKADIROGLU, A., J. ANGRIST, S. DYNARSKI, T. KANE, AND P. PATHAK (2011): "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots," *Quarterly Journal of Economics*, 126(2), 699–748.

ABDULKADIROGLU, A., J. ANGRIST, AND P. PATHAK (2014): "The Elite Illusion: Achievement Effects at Boston and New York Exam Schools," *Econometrica*, 82(1), 137–196.

AIGNER, D., C. HSIAO, A. KAPTEYN, AND T. WANSBEEK (1984): "Latent Variable Models in Econometrics," Handbook of Econometrics 2, pp. 1321–1393. Elsevier B.V., Amsterdam.

AIT-SAHALIA, Y., P. BICKEL, AND T. STOKER (2001): "Goodness-of-Fit Tests for Kernel Regression with an Application to Option Implied Volatilities," *Journal of Econometrics*, 105(2), 363–412.

ALMOND, D., J. DOYLE, A. KOWALSKI, AND H. WILLIAMS (2010): "Estimating Marginal Returns to Medical Care: Evidence from At-risk Newborns," *Quaterly Journal of Economics*, 125(2), 591–634.

ANDREWS, D. (2002): "Higher-Order Improvements of a Computationally Attractive k-Step Bootstrap for Extremum Estimators," *Econometrica*, 70(1), 119–162.

——— (2011): "Examples of L2-Complete and Boundedly-Complete Distributions," Cowles Foundation Discussion Paper 1801.

ANDREWS, D., AND P. GUGGENBERGER (2009): "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77(3), 721–762.

ANDREWS, R., V. RANCHHOD, AND V. SATHY (2010): "Estimating the Responsiveness of College Applications to the Likelihood of Acceptance and Financial Assistance: Evidence from Texas," *Economics of Education Review*, 29(1), 104–115.

ANGRIST, J., S. COHODES, S. DYNARSKI, P. PATHAK, AND C. WALTERS (2013): "Charter Schools and the Road to College Readiness: The Effects on College Preparation, Attendance and Choice," Research Report Prepared for the Boston Foundation and NewSchools Venture Fund.

ANGRIST, J., AND I. FERNANDEZ-VAL (2010): "ExtrapoLATE-ing: External Validity and Overidentification in the LATE Framework," NBER Working Paper 16566.

ANGRIST, J., AND G. IMBENS (1995): "Two-Stage Least Squares Estimation of Average Causal Effects in Models with Variable Treatment Intensity," *Journal of the American Statistical Association*, 90(430), 431 442.

ANGRIST, J., G. IMBENS, AND D. RUBIN (1996): "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91(434), 444 455.

ANGRIST, J., AND G. KUERSTEINER (2011): "Causal Effects of Monetary Shocks: Semiparametric Conditional Independence Tests with a Multinomial Propensity Score," *Review of Economics and Statistics*, 93(3), 725–747.

ANGRIST, J., AND J.-S. PISCHKE (2009): *Mostly Harmless Econometrics*. Princeton University Press, Princeton.

ANGRIST, J., AND M. ROKKANEN (2013): "Wanna Get Away? RD Identification Away from the Cutoff," IZA Discussion Paper 7429.

ANGRIST, J. D., AND V. LAVY (1999): "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quaterly Journal of Economics*, 114(2), 533–575.

ANTONOVICS, K., AND B. BACKES (2013): "Were Minority Students Discouraged from Applying to University of California Campuses after the Affirmative Action Ban?," *Education Finance and Policy*, 8(2), 208–250.

——— (forthcoming): "The Effect of Banning Affirmative Action on College Admissions Policies and Student Quality," *Journal of Human Resources*.

ARAI, Y., AND H. ICHIMURA (2013): "Bandwidth Selection for Differences of Nonparametric Estimators with an Application to the Regression Discontinuity Design," Unpublished manuscript.

ARCIDIACONO, P. (2005): "Affirmative Action in Higher Education: How Do Admission and Financial Aid Rules Affect Future Earnings?," *Econometrica*, 73(5), 1477–1524.

BAR-LEV, S., AND D. PLACHKY (1989): "Boundedly Complete Families Which Are Not Complete," *Metrika*, 36(1), 331–336.

BARGAIN, O., AND K. DOORLEY (2013): "Putting Structure on the RD Design: Social Transfers and Youth Inactivity in France," IZA Discussion Paper 7508.

BARNOW, B. (1972): *Conditions for the Presence or Absence of a Bias in Treatment Effect: Some Statistical Models for Head Start Evaluation*. University of Wisconsin-Madison.

BATTISTIN, E., C. LAMARCHE, AND E. RETTORE (2013): "Latent Structures and Quantiles of the Treatment Effect Distribution," Unpublished manuscript.

BATTISTIN, E., AND E. RETTORE (2008): "Ineligibles and Eligible Non-Participants as a Double Comparison Group in Regression-Discontinuity Designs," *Journal of Econometrics*, 142(2), 715–730.

BELLONI, A., V. CHERNOZHUKOV, AND C. HANSEN (2012): "Inference on Treatment Effects After Selection Amongst High-Dimensional Controls," ArXiv:1201.02243.

BERTRAND, M., R. HANNA, AND S. MULLAINATHAN (2010): "Affirmative Action in Education: Evidence from Engineering College Admissions in India," *Journal of Public Economics*, 94(1-2), 16- 29.

BHARADWAJ, P., K. LOKEN, AND C. NEILSON (2013): "Early Life Health Interventions and Academic Achievement," *American Economic Review*, 103(5), 1862–1891.

BLACK, S. E. (1999): "Do Better Schools Matter? Parental Valuation of Elementary Education," *Quaterly Journal of Economics*, 114(2), 577–599.

BLOOM, H. (2012): "Modern Regression Discontinuity Analysis," *Journal of Research on Educational Effectiveness*, 5(1), 43–82.

BLUNDELL, R., X. CHEN, AND D. KRISTENSEN (2007): "Semi-Nonparametric IV Estimation of Shape-Invariant Engel Curves," *Econometrica*, 75(6), 1613–1669.

BLUNDELL, R., AND J. POWELL (2003): "Endogeneity in Nonparametric and Semiparametric Regression Models," Advances in Economics and Econometrics: Theory and Applications 2, pp. 312–357. Cambridge University Press, Cambridge.

BONHOMME, S., AND U. SAUDER (2011): "Recovering Distributions in Difference-in-Differences Models: A Comparison of Selective and Comprehensive Schooling," *Review of Economic Studies*, 93(2), 479–494.

BOUND, J., C. BROWN, AND N. MATHIOWETZ (2001): "Measurement Error in Survey Data," Handbook of Econometrics 5, pp. 3705–3843. Elsevier B.V., Amsterdam.

Bui, S., S. Craig, and S. Imberman (forthcoming): "Is Gifted Education a Bright Idea? Assessing the Impacts of Gifted and Talented Programs," *American Economic Journal: Economic Policy*.

Calonico, S., M. Cattaneo, and R. Titiunik (2014): "Robust Nonparametric Confidence Intervals for Regression-Discontinuity Designs," Unpublished manuscript.

Campbell, D., and J. Stanley (1963): *Experimental and Quasi-Experimental Designs for Research*. Houghton Mifflin Company, Boston.

Canay, I., A. Santos, and A. Shaikh (forthcoming): "On the Testability of Identification in Some Nonparametric Models with Endogeneity," *Econometrica*.

Card, D., and A. Krueger (2005): "Would the Elimination of Affirmative Action Affect Highly Qualified Minority Applicants? Evidence from California and Texas," *Industrial & Labor Relations Review*, 58(3), 416–434.

Card, D., D. S. Lee, Z. Pei, and A. Weber (2012): "Nonlinear Policy Rules and the Identification and Estimation of Causal Effects in a Generalized Regression Kink Design," Unpublished manuscript.

Carneiro, P., K. Hansen, and J. Heckman (2001): "Removing the Veil of Ignorance in Assessing the Distributional Impacts of Social Policiese," *Swedish Economic Policy Review*, 8(2), 273–301.

——— (2003): "Estimating Distributions of Counterfactuals with an Application to the Returns to Schooling and Measurement of the Effect of Uncertainty on Schooling Choice," *International Economic Review*, 44(2), 361 422.

Carroll, R., D. Ruppert, L. Stefanski, and C. Crainiceanu (2006): *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman & Hall/CRC, Boca Raton.

Chen, X., V. Chernozhukov, S. Lee, and W. Newey (2013): "Local Identification of Nonparametric and Semiparametric Models," Unpublished manuscript.

Chen, X., H. Hong, and D. Nekipelov (2011): "Nonlinear Models of Measurement Errors," *Journal of Economic Literature*, 49(4), 901–937.

Cheng, M.-Y., J. Fan, and J. S. Marron (1997): "On Automatic Boundary Corrections," *The Annals of Statistics*, 25(4), 1691–1708.

Chernozhukov, V., and C. Hansen (2005): "An IV Model of Quantile Treatment Effects," *Econometrica*, 73(1), 245–261.

Chernozhukov, V., G. Imbens, and W. Newey (2007): "Instrumental Variable Estimation of Nonseparable Models," *Journal of Econometrics*, 139(1), 4–14.

COOK, T. (2008): ""Waiting for Life to Arrive": A History of the Regression-Disconinuity Design in Psychology, Statistics and Economics," *Journal of Econometrics*, 142(2), 636–654.

COOK, T., AND C. WING (2013): "Strengthening The Regression Discontinuity Design Using Additional Design Elements: A Within-Study Comparison," Unpublished manuscript.

COOLEY FRUEHWIRTH, J., S. NAVARRO, AND Y. TAKAHASHI (2011): "How the Timing of Grade Retention Affects Outcomes: Identification and Estimation of Time-Varying Treatment Effects," Unpublished manuscript.

CORTES, K. (2010): "Do Bans on Affirmative Action Hurt Minority Students? Evidence from the Texas Top 10% Plan," *Economics of Education Review*, 29(6), 1110–1124.

CUNHA, F., J. HECKMAN, AND S. NAVARRO (2005): "Separating Uncertainty from Heterogeneity in Life Cycle Earnings," *Oxford Economic Papers*, 57(2), 191–261.

CUNHA, F., J. HECKMAN, AND S. SCHENNACH (2010): "Estimating the Technology of Cognitive and Noncognitive Skill Formulation," *Econometrica*, 78(3), 883–931.

DAROLLES, S., Y. FAN, J.-P. FLORENS, AND E. RENAULT (2011): "Nonparametric Instrumental Regression," *Econometrica*, 79(5), 1541–1565.

DAVIDSON, R., AND J. MACKINNON (1999): "Bootstrap Testing in Nonlinear Models," *International Economic Review*, 40(2), 487–508.

DEHEJIA, R., AND S. WAHBA (1999): "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *Journal of the American Statistical Association*, 94(448), 1053–1062.

DESJARDINS, S. L., AND B. P. MCCALL (2008): "The Impact of the Gates Millennium Scholars Program on the Retention, College Finance- and Work-Related Choices, and Future Educational Aspirations of Low-Income Minority Students," Unpublished manuscript.

D'HAULTFOEUILLE, X. (2011): "On the Completeness Condition in Nonparametric Instrumental Problems," *Econometric Theory*, 27(3), 460–471.

DICKSON, L. (2006): "Does Ending Affirmative Action in College Admissions Lower the Percent of Minority Students Applying to College?," *Economics of Education Review*, 25(1), 109–119.

DINARDO, J., AND D. LEE (2011): "Program Evaluation and Research Designs," Handbook of Labor Economics 4A, pp. 465–536. Elsevier B.V., Amsterdam.

DOBBIE, W., AND R. FRYER (2013): "Exam High Schools and Student Achievement: Evidence from New York City," Unpublished manuscript.

DOKSUM, K., D. PETERSON, AND A. SAMAROV (2000): "On Variable Bandwidth Selection in Local Polynomial Regression," *Journal of the Royal Statistical Society: Series B*, 62(3), 431–448.

DONG, Y., AND A. LEWBEL (2013): "Identifying the Effect of Changing the Policy Threshold in Regression Discontinuity Models," Unpublished manuscript.

DUFLO, E., P. DUPAS, AND M. KREMER (2011): "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randimized Evaluation in Kenya," *American Economic Review*, 101(5), 1739–1774.

EPPLE, D., AND R. ROMANO (2011): "Peer Effects in Education: A Survey of the Theory and Evidence," Handbook of Social Economics 1B, pp. 1053–1163. Elsevier B.V., Amsterdam.

EVDOKIMOV, K., AND H. WHITE (2012): "Some Extensions of a Lemma of Kotlarski," *Econometric Theory*, 28(4), 925–932.

FAN, J., T. GASSER, I. GIJBELS, M. BROCKMANN, AND J. ENGEL (1997): "Local Polynomial Regression: Optimal Kernels and Asymptotic Minimax Efficiency," *Annals of the Institute of Statistical Mathematics*, 49(1), 79–99.

FAN, J., AND I. GIJBELS (1992): "Variable Bandwidth and Local Linear Regression Smoothers," *The Annals of Statistics*, 20(4), 2008–2036.

——— (1996): *Local Polynomial Modelling and Its Applications*. Chapman & Hall, London.

FAN, Y., AND Q. LI (1996): "Consistent Model Specification Tests: Omitted Variables and Semiparametric Functional Forms," *Econometrica*, 64(4), 865–890.

FLORENS, J.-P., M. MOUCHART, AND J.-M. ROLIN (1990): *Elements of Bayesian Statistics*. Marcel Dekker, New York.

FRANCIS, A., AND M. TANNURI-PIANTO (2012): "Using Brazil's Racial Continuum to Examine the Short-Term Effects of Affirmative Action in Higher Education," *Journal of Human Resources*, 47(3), 754–784.

FRANDSEN, B., M. FROLICH, AND B. MELLY (2012): "Quantile Treatment Effects in the Regression Discontinuity Design," *Journal of Econometrics*, 168(2), 382–395.

FROLICH, M. (2007): "Nonparametric IV Estimation of Local Average Treatment Effects with Covariates," *Journal of Econometrics*, 139(1), 35–75.

GALE, D., AND L. SHAPLEY (1962): "College Admissions and the Stability of Marriage," *American Mathematical Monthly*, 69(1), 9–15.

GERARD, P. D., AND W. R. SCHUCANY (1997): "Methodology for Nonparametric Regression from Independent Sources," *Computational Statistics & Data Analysis*, 25, 287–304.

GHOSH, J., AND R. SINGH (1966): "Unbiased Estimation of Location and Scale Parameters," *Annals of Mathematical Statistics*, 37(6), 1671–1675.

GOLDBERGER, A. (1972a): "Selection Bias in Evaluating Treatment Effects: Some Formal Illustrations," Unpublished manuscript.

——— (1972b): "Selection Bias in Evaluating Treatment Effects: the Case of Interaction," Unpublished manuscript.

GRAHAM, B. (2011): "Econometric Methods for the Analysis of Assignment Problems in the Presence of Complementarity and Social Spillovers," Handbook of Social Economics 1B, pp. 965–1052. Elsevier B.V., Amsterdam.

GRAHAM, B., G. IMBENS, AND G. RIDDER (2010): "Measuring the Effects of Segregation in the Presence of Social Spillovers: A Nonparametric Approach," NBER Working Paper 16499.

——— (2013): "Complementarity and Aggragate Implications of Assortative Matching: A Nonparametric Analysis," Unpublished manuscript.

HAHN, J., P. TODD, AND W. VAN DER KLAAUW (2001): "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design," *Econometrica*, 69(1), 201–209.

HALL, P., AND J. HOROWITZ (2005): "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *Annals of Statistics*, 33(6), 2904–2929.

HANSEN, L. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.

HAUSMAN, J. (2001): "Mismeasured Variables in Econometric Analysis: Problems from the Right and Problems from the Left," *Journal of Economic Perspectives*, 15(4), 57–67.

HECKMAN, J., H. ICHIMURA, AND P. TODD (1998): "Matching as an Econometric Evaluation Estimator," *Review of Economic Studies*, 65(2), 261–294.

HINRICHS, P. (2012): "The Effects of Affirmative Action Bans on College Enrollment, Educational Attainment, and the Demographic Composition of Universities," *Review of Economics and Statistics*, 94(3), 712–722.

HIRANO, K., G. IMBENS, AND G. RIDDER (2003): "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4), 1161–1189.

HOEFFDING, W. (1977): "Some Incomplete and Boundedly Complete Families of Distributions," *Annals of Statistics*, 5(5), 278–291.

HONG, H., AND D. NEKIPELOV (2010): "Semiparametric Efficiency in Nonlinear LATE Models," *Quantitative Economics*, 1(2), 279–304.

HOROWITZ, J. (2011): "Applied Nonparametric Instrumental Variables Estimation," *Econometrica*, 79(2), 347–394.

HORVITZ, D., AND D. THOMPSON (1952): "A Generalization of Sampling Without Replacement From a Finite Universe," *Journal of the American Statistical Association*, 47(260), 663–685.

HOWELL, J. (2010): "Assessing the Impact of Eliminating Affirmative Action in Higher Education," *Journal of Labor Economics*, 28(1), 113–166.

HU, Y., AND S. SCHENNACH (2008): "Instrumental Variable Treatment of Nonclassical Measurement Error Models," *Econometrica*, 76(1), 195–216.

HU, Y., AND J.-L. SHIU (2012): "Nonparametric Identification using Instrumental Variables: Sufficient Conditions for Completeness," Unpublished manuscript.

IMBENS, G. (2004): "Nonparametric Estimation of Average Treatment Effects under Exogeneity: A Review," *Review of Economic Studies*, 86(1), 4 29.

IMBENS, G., AND J. ANGRIST (1994): "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62(2), 467–475.

IMBENS, G., AND K. KALYANARAMAN (2012): "Optimal Bandwidth Choice for the Regression Discontinuity Estimator," *Review of Economic Studies*, 79(3), 933 959.

IMBENS, G., AND T. LEMIEUX (2008): "Regression Discontinuity Designs: A Guide to Practice," *Journal of Econometrics*, 142(2), 615–635.

IMBENS, G. W., AND J. M. WOOLDRIDGE (2009): "Recent Developments in the Econometrics of Program Evaluation," *Journal of Economic Literature*, 47(1), 615–635.

ISENBECK, M., AND L. RUSCHENDORF (1992): "Completeness in Location Families," *Probability and Mathematical Statistics*, 13(2), 321–343.

JACKSON, K. (2009): "Student Demographics, Teacher Sorting, and Teacher Quality: Evidence from the End of School Desegregation," *Journal of Labor Economics*, 27(2), 213–256.

——— (2010): "Do Students Benefit from Attending Better Schools? Evidence from Rule-Based Student Assignments in Trinidad and Tobago," *Economic Journal*, 120(549), 1399–1429.

JACOB, B., AND L. LEFGREN (2004): "Remedial Education and Student Achievement: A Regression-Discontinuity Analysis," *Review of Economic and Statistics*, 86(1), 226–244.

KLINE, P. (2011): "Oaxaca-Blinder as a Reweighting Estimator," *American Economic Review: Papers and Proceedings*, 101(3), 532–537.

KOTLARSKI, I. (1967): "On Characterizing the Gamma and the Normal Distribution," *Pacific Journal of Mathematics*, 20(1), 69–76.

LALONDE, R. (1986): "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *American Economic Review*, 76(4), 604–620.

LAVERGNE, P., AND Q. VUONG (2000): "Nonparametric Significance Testing," *Econometric Theory*, 16(4), 576–601.

LEE, D. (2008): "Randomized Experiments from Non-Random Selection in U.S. House Elections," *Journal of Econometrics*, 142(2), 675–697.

LEE, D., AND T. LEMIEUX (2010): "Regression Discontinuity Designs in Economics," *Journal of Economic Literature*, 48, 281–355.

LEHMANN, E., AND J. ROMANO (2005): *Testing Statistical Hypothesis*. Springer, New York.

LEHMANN, E., AND H. SCHEFFE (1950): "Completeness, Similar Regions, and Unbiased Estimation: Part I," *Sankhya*, 10(4), 305–340.

——— (1955): "Completeness, Similar Regions, and Unbiased Estimation: Part II," *Sankhya*, 15(3), 219–236.

LEWBEL, A. (2007): "Estimation of Average Treatment Effects with Misclassification," *Econometrica*, 2(3), 537–551.

LI, K.-C. (1987): "Asymptotic Optimality for Cp, CL, Cross-Validation and Generalized Cross-Validation: Discrete Index Set," *The Annals of Statistics*, 15(3), 958–975.

LONG, M. (2004): "College Applications and the Effects of Affirmative Action," *Journal of Econometrics*, 121(1-2), 319–342.

LORD, F., AND M. NOVICK (1972): *Statistical Theories of Mental Test Scores*. Addison-Welsley.

LUDWIG, J., AND D. L. MILLER (2005): "Does Head Start Improve Children's Life Changes? Evidence from a Regression Discontinuity Design," NBER Working Paper 11702.

——— (2007): "Does Head Start Improve Children's Life Changes? Evidence from a Regression Discontinuity Design," *Quaterly Journal of Statistics*, 122(1), 159 -208.

MATSUDAIRA, J. (2008): "Mandatory Summer School and Student Achievement," *Journal of Econometrics*, 142(8), 829–850.

MATTNER, L. (1992): "Completeness of Location Families, Translated Moments, and Uniqueness of Charges," *Probability Theory and Related Fields*, 92(2), 137–149.

——— (1993): "Some Incomplete But Boundedly Complete Location Families," *Annals of Statistics*, 21(4), 2158–2162.

NEWEY, W., AND J. POWELL (2003): "Instrumental Variable Estimation of Nonparametric Models," *Econometrica*, 71(5), 1565–1578.

PORTER, J. (2003): "Estimation in the Regression Discontinuity Model," Unpublished manuscript.

PRAKASA RAO, B. (1992): *Identifiability in Stochastic Models: Characterization of Probability Distributions*. Academic Press, Boston.

PREWITT, K. A. (2003): "Efficient Bandwidth Selection in Non-Parametric Regression," *Scandinavian Journal of Statistics*, 30(1), 75–92.

ROKKANEN, M. (2014): "Exam Schools, Ability, and the Effects of Affirmative Action: Latent Factor Extrapolation in the Regression Discontinuity Design," Unpublished manuscript.

ROTHSTEIN, J., AND A. YOON (2008): "Affirmative Action in Law School Admissions: What Do Racial Preferences Do?," *University of Chicago Law Review*, 75(2), 649–714.

RUPPERT, D. (1997): "Empirical-Bias Bandwidths for Local Polynomial Nonparametric Regression and Density Estimation," *Journal of American Statistical Association*, 92(439), 1049–1062.

SACERDOTE, B. (2011): "Peer Effects in Education: How Might They Work, How Big Are They and How Much Do We Know Thus Far?," Handbook of Economics of Education 3, pp. 249–277. Elsevier B.V., Amsterdam.

SANTOS, A. (2012): "Inference in Nonparametric Instrumental Variables with Partial Identification," *Econometrica*, 80(1), 213–275.

SCHUCANY, W. R. (1995): "Adaptive Bandwidth Choice for Kernel Regression," *Journal of the American Statistical Association*, 90(430), 535–540.

SEVERINI, T., AND G. TRIPATHI (2006): "Some Identification Issues in Nonparametric Linear Models with Endogenous Regressors," *Econometric Theory*, 22(2), 258–278.

STONE, C. J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression," *The Annals of Statistics*, 10(4), 1040–1053.

THISTLEWHITE, D. L., AND D. T. CAMPBELL (1960): "Regression-Discontinuity Analysis: An Alternative to the Ex Post Facto Experiment," *Journal of Educational Psychology*, 51(6), 309–317.

VAN DER KLAAUW, W. (2002): "Estimating the Effect of Financial Aid Offers on College Enrollment - A Regression-Discontinuity Approach," *International Economic Review*, 43(4), 1249–1287.

——— (2008): "Regression-Discontinuity Analysis: A Survey of Recent Developments in Economics," *Labour*, 22(2), 219–245.

WALTERS, C. (2013): "A Structural Model of Charter School Choice and Academic Achievement," Unpublished manuscript.