NBER WORKING PAPER SERIES

EVALUATING THE GIFTED PROGRAM OF AN URBAN SCHOOL DISTRICT USING A MODIFIED REGRESSION DISCONTINUITY DESIGN

Billie Davis
John Engberg
Dennis N. Epple
Holger Sieg
Ron Zimmer

Working Paper 16414
http://www.nber.org/papers/w16414

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
September 2010

We would like to thank William Dickens, Guido Imbens, Brian Junker, Allen Ruby, Robert Siegler, and seminar participants at numerous conferences and universities for comments and suggestions. We would also like to thank the "mid-sized urban school district" for sharing their data. Andrea Phillips provided excellent research assistance. Financial support for this research is provided by the Institute of Education Sciences (IES R305A070117 and R305D090016). The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Evaluating the Gifted Program of an Urban School District using a Modified Regression Discontinuity Design

Billie Davis, John Engberg, Dennis N. Epple, Holger Sieg, and Ron Zimmer

NBER Working Paper No. 16414
September 2010
JEL No. I21

ABSTRACT

This paper examines the impact of a gifted program on retention in an urban school district using a regression discontinuity design. Gifted programs often employ IQ thresholds for admission, with those above the threshold being admitted. One common problem with the RD design arises if the forcing variable (the IQ score) is manipulated, thus invalidating the standard research design. We proposed a modified RD estimator that deals with manipulation in the forcing variable. Once we properly correct for manipulation of test scores around the cut-off points, we find evidence that the gifted program offered by the district has a positive effect on retention of higher income students.

Billie Davis
Carnegie Mellon University
Tepper School of Business
5000 Forbes Ave
Pittsburgh, PA 15213
billie@cmu.edu

John Engberg
Rand Corporation
4570 5th Avenue
Pittsburgh, PA 15213
engberg@rand.org

Dennis N. Epple
Tepper School of Business
Carnegie Mellon University
Posner Hall, Room 257B
Pittsburgh, PA 15213
and NBER
epple@cmu.edu

Holger Sieg
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104
and NBER
holgers@econ.upenn.edu

Ron Zimmer
Erickson Hall
Michigan State University
East Lansing, MI 48824
rzimmer@msu.edu

1 Introduction

Gifted and talented programs have a long history in the U.S., dating back to the late 19^{th} century. However, gifted programs did not receive federal support until 1958 when the federal government established the National Defense Education Act. This act initiated federal support for specialized programs for math, science, and foreign languages (Bhatt, 2009). More recently, the federal government expanded its support to gifted programs through the Jacob Javits Gifted and Talented Educational Act in 1988 and the No Child Left Behind Act in 2002. Through these initiatives, gifted programs have gained popularity, especially in urban districts.

For urban districts, these programs have the dual objective of engaging and challenging gifted students to reach advanced levels of achievement as well as attracting and retaining students who might otherwise leave for suburban or private schools. Despite receiving federal support, gifted programs are not mandated by the federal government. Individual states or districts decide if and how to use gifted programs, including how students are identified (Shaunessy, 2003). While there is much interest in gifted education outside of economics, few economists have directly addressed this topic.[1]

The first purpose of this paper is to evaluate the admission procedures and

[1] A meta-analysis of nine papers by Vaughn, Feldhusen, and Asher (1991) found that participation in pull-out gifted programs led to improved achievement, critical thinking, and creativity, but student's self-concepts were not affected. However, these studies often had difficulties dealing with thorny endogeneity issues and were not all peer-reviewed. Bhatt (2009) finds that participation in gifted programs leads to strong initial effects in math that dissipate over time and an increase in the long run in the probability that a child takes Advanced Placement classes. However, the results provide little evidence that gifted education increases interest and participation in academics or affects a student's peer group.

to estimate the treatment effect of admittance into a gifted program on retention in a mid-size urban school district. Student retention has increasingly become an important issue for large urban districts, especially in the Midwest and East Coast. Districts such as Buffalo, Cincinnati, Cleveland, Detroit, Kansas City, Milwaukee, Pittsburgh, and Philadelphia have lost thousands of students over the last several decades. In many cases, these districts' enrollments have been cut in half over the last 40 to 50 years (Zimmer, Guarino and Buddin, 2010). State funding, which is allocated on a per-pupil basis, has therefore shrunk dramatically. The declining urban population has also led to a lower local tax base. However, these districts have buildings, staffing, and pension systems designed for a much larger enrollment base. Together, the pressure both on the cost and revenue side have caused districts to search for ways to help retain students, including gifted programs. In searching for a school, families may consider not only the quality of facilities, curriculum, and instruction, but also consider the quality of educational opportunities and peers. Gifted programs may create opportunities for students to be stimulated and challenged and have positive peer influences. Often, smaller suburban districts may not have the scale to offer such programs and as a result, gifted programs may be a mechanism for retaining strong students within the district.

Many gifted programs, including the one studied in this paper, employ IQ thresholds for admission, with those above the threshold being admitted. Psychologists administering IQ tests may "give the benefit of the doubt" in assessing performance of students who are near the threshold for admission to a gifted program. These types of activities imply that the observed IQ scores are at best noisy measures of the underlying ability of the students. In many cases, the test scores appear to be manipulated.[2] One consequence of this manipulation is that

[2]The IQ tests are oral rather than written, with the examiner scoring the students responses.

IQ scores at the cut-off point cannot be used as instruments for program participation, which invalidates the key identifying assumption of the standard regression discontinuity design.[3] The second purpose of the paper is then to show how to deal with manipulation in the forcing variable in a RD design.

A distinctive feature of educational environments is that sub-scores and alternative composite scores are often available and may be considered along with the overall IQ score in determining admission. Some of these sub-scores or composites are directly referenced in admission guidelines and are therefore also likely to be manipulated. Other sub-scores are not directly referenced and are therefore less likely to be manipulated. It is likely that even though all sub-scores are reported to parents, they would pay most attention to the scores that are directly referenced in the admission guidelines.

As a consequence, it is plausible to assume that the sub-tests not directly referenced are uncorrelated with unobservable components in parental decisions such as retention. However, they are clearly correlated with the overall IQ score used to determine access to the program.

We can use these sub-scores to predict an IQ. The basic idea behind our modified regression discontinuity estimator is then to use the exogenous part of the discontinuous change in program participation at the threshold that is predicted by the sub-scores as an instrument for program participation. We discuss the conditions under which such additional testing information can overcome the dif-

Hence, there is some scope for exercise of discretion by the examiner.

[3]Similar issues arise in other educational settings. Concerns for transparency typically lead to promulgation of the criteria for admission. Knowing the criteria for admission, participants may then undertake activities that alter the reported outcome on the variable that determines admission. For example, students who fall below the threshold on a test determining whether they will be subject to remediation may retake the test to attempt to obtain a score above the threshold (Calcagno and Long, 2008).

ficulties that the manipulation of the key forcing variable creates for the regression discontinuity design.[4]

Our application focuses on a gifted program operated by a mid-sized urban school district that prefers not to be identified. We implement our estimation strategy for a sample of students tested for the gifted program while attending a district school in school years 2003 to 2007. Our findings suggest that our modified RD approach works well in this application. Using the predicted scores yields plausible estimates of the admission probabilities around the discontinuity. The relatively large positive point estimates suggest that there is a favorable effect on retention for higher income students (i.e., those not on subsidized lunch). The estimated standard errors are also relatively large, however, so these effects are imprecisely estimated.

Our research is closely related to two well-established lines of literature – on tracking and peer effects. Tracking is the practice of assigning students to classes based on the ability of students. Previous research has focused on the effect tracking programs have on test scores.[5]. Those who advocate for tracking argue that a teacher can tailor the curriculum to the ability level of the students, thus creating the optimal level of educational gains for all students. Opponents argue against tracking for three primary reasons. First, tracking leads to a different set of resources being allocated to high-tracked versus low-tracked classes (Oakes, 1990).

[4]Since the late 1990s there have been a number of studies in educational economics and education science that apply regression discontinuity design (RD) methods. The RD design was first used by Thistlethwaite and D.Cambell (1960). Recent applications of the RD design in education include Angrist and Lavy (1999), van der Klaauw (2002), Jacob and Lefgren (2004), and Chay, McEwan, and Urquiola (2005).

[5]See, for example,Duflo, Dupas, and Kremer (2008), Zimmer (2003), Figlio and Page (0002), Argys, Rees, and Brewer (1996), Hoffer (1992), and Kerckhoff (1986). See Epple, Newlon, and Romano (2002) for a theoretical analysis

Second, tracking breeds social inequities as minority and low-income groups are over-represented in low-track and under-represented in high-track classes.[6]. Third, tracking creates homogeneous classes according to ability, therefore reducing the positive spillover effect, referred to as a peer effect (Betts and Shkolnik, 2000). Gifted programs, however, are often pull-out programs in which students spend time outside of their regular classroom to gain specialized instruction and therefore do not always fit into the tracking framework. While ability grouping and peer effects may contribute to the effects of a gifted program on test scores, our focus is on examining the impact that gifted programs have on student retention.

The rest of the paper is organized as follows. Section 2 provides information about our data set and describes the testing and admission procedures used in the gifted program studied in our application. Section 3 develops the modified regression discontinuity design estimator that accounts for manipulation in the criterion variable. Section 4 presents the empirical findings of our paper. It documents the extent of manipulation of IQ scores and discusses the role that additional sub-scores play in the first stage of the RDD estimator. We then study the effectiveness of the gifted program in retaining students in district schools. Section 5 offers some conclusions and discusses future research.

2 Data

2.1 Institutional Background

The school district that we study in this paper operates a gifted program that is quite large in scope. Gifted students in grades 1 to 8 participate in a one-day-per-week pull out program at a designated location away from the student's home

[6]See Braddock and Dawkins (1993), Gorman (1987), Oakes (1990).

school. Students enroll in programs designed to enhance creative problem solving and leadership skills and are offered specially designed instruction in math, science, literature, and a variety of other fields. For high school students, gifted education is available within the school and involves the annual design of an individualized education program, full-time curricula, and a number of other enhancements.

The district adheres to state regulations concerning gifted students and services. The state regulations outline a multifaceted approach used to identify whether a student is gifted and whether gifted education is needed. A mentally gifted student is defined as someone with an IQ of at least 130 points or someone who shows outstanding intellectual and creative ability using other educational criteria. Further, to qualify for gifted services, the district must show that the student requires services or programs not available in regular education.

The state guidelines stress that IQ cannot be the only factor used in determining gifted ability. Specifically, low scores in memory or processing speed tests cannot be used alone to disqualify a student. Also, even if a student has an IQ below 130, she may be deemed gifted based on above grade level achievement on standardized tests, a superior rate of acquisition or retention of new academic content or skills, excellence in specific academic areas, or other factors that indicate superior functioning. Additionally, the guidelines specifically note that the gifted decision must account for any potential masking of gifted abilities because of disability, socio/cultural deprivation, gender or race bias, or English as a second language. Further, it is emphasized that the gifted decision may not be based on a single test or type of test. For limited English proficiency or students of racial-, linguistic-, or ethnic-minority background, it is specifically noted that an IQ score may not be used as the only measure to show low aptitude.

The evaluation process begins when a parent, teacher, administrator, or student requests a gifted evaluation. Once the student's parents are notified and give

consent for the evaluation, a team consisting of parents, a certified school psychologist, teachers, and others familiar with the student's educational experience and performance or cultural background conducts the evaluation. The evaluation must include information on academic functioning, learning strengths, and educational needs. The information and findings from the evaluation of the student's educational needs and strengths is combined by the team into a written report. This report includes the team's recommendation as to whether the student is gifted and in need of specially designed instruction. Finally, the report is evaluated in a team meeting where the decision is made regarding the student's eligibility for gifted education.

The district adheres to the preceding guidelines for evaluating potential gifted students. As noted, one way to support a claim of giftedness is to show superior performance (above 130 points) on an intelligence test. In our district, every student considered for the gifted program is given some type of intelligence test. During the time-frame of our analysis (school years 2004-2005 to 2007-2008) district psychologists mainly used the Wechsler Intelligence Scale for Children, 4^{th} edition (WISC4) test instrument.[7] The WISC4 gives four index scores measur-

[7]The WISC was updated from the 3^{rd} edition to the 4^{th} edition in 2003. During the 2004-2005 school year, the district phased out use of the WISC3 and phased in use of the WISC4. By the 2005-2006 school year, the WISC3 was no longer used by district psychologists. The WISC4 test instrument is designed for students 6 years to 16 years 11 months old. For younger students, the Wechsler Preschool and Primary Scale of Intelligence, 3^{rd} edition (WPPSI-III) (ages 2 years 6 months to 7 years 3 months) was mainly used. The Stanford Binet Intelligence Scale, 5^{th} edition (SB-V) (ages 2 to 85 years) was also sometimes used for younger students. For older students, psychologists used the Wechsler Adult Intelligence Scale, 3^{rd} edition (WAIS-III) (ages 16 to 89) or the Woodcock-Johnson III Tests of Cognitive Abilities (WJ-III) (ages 2 to 90+ years).

Additionally, for culturally or linguistically diverse students, a comprehensive non-verbal measure may have been used in place of or alongside the scores from the above tests. Acceptable

ing verbal comprehension (VCI), perceptual reasoning (PRI), working memory (WMI), and processing speed (PSI). It also gives a "Full Scale IQ" (FSIQ) which combines the results from the four indexes and a "Generalized Ability Index" (GAI) which combines the results from the VCI and PRI. The FSIQ, indexes, and GAI are normed, by age, to be representative of the current population of children in the United State and have a mean of 100 and a standard deviation of 15. Thus, a score of 130 is two standard deviations above the mean.[8]

In the district, each student takes an intelligence test and is then categorized as meeting the IQ criteria if the FSIQ or GAI is 130 or above.[9] Students with a FSIQ or GAI of 125 to 129 or a VCI or PRI of 130 or above do not meet the IQ criteria but do qualify for special further consideration through a "portfolio evaluation". Additionally, students eligible for subsidized lunch with an FSIQ or GAI of 115 to 124 also receive a portfolio evaluation. Students who score below these cutoffs may still be considered gifted based on a further review of other factors. In practice, the probability of a student being admitted into the gifted

non-verbal tests or tests with non-verbal measures included the Naglieri Non-Verbal Ability Test, Individual edition (NNAT-I), the Comprehensive Test of Nonverbal Intelligence (C-TONI), the Universal Nonverbal Intelligence Test (UNIT), the Leiter International Performance Scale, Revised edition (Leiter-R), the Kaufman Assessment Battery for Children, 2^{nd} edition (K-ABC-II), the Differential Ability Scales (DAS), and the SB-V.

[8]For more information on the WISC4 test instrument, see the publisher's web page at http://www.pearsonassessments.com. The various testing instruments (and editions thereof) measure different aspects of functioning, report different test composites, are separately normed, and may focus on different subsets of the population. Therefore, one must use caution in comparing results from one test to another. Since the district uses the WISC4 unless the student is too young, too old, or is culturally or linguistically diverse, we focus on students who were given the WISC4 test.

[9]Note that the GAI excludes processing speed and working memory sub-tests. Hence, the GAI offers a way to address the state requirement pertaining to not excluding students based solely on low scores in memory and processing speed.

program increases most at the portfolio cutoff of 125 points for regular lunch students and 115 points for subsidized lunch students.

2.2 The Sample

We have student level data for all students attending district schools in school years 2004 to 2007 including information about race, gender, standardized test scores, subsidized lunch status, school attended, home census tract, etc. Tables 1 and 2 summarize a variety of student characteristics. Column 2 of these tables reports averages for students in kindergarten through 12^{th} grade enrolled in the district at some time between 2004 and 2007, with standard deviations in parentheses.

In Table 1 we see that 51% of district students are male, 56% are African American, and 77% receive subsidized lunch.[10] For the students' census tracts, the average median income is $28,868 and 18.5% of adults have earned at least a bachelor's degree. Table Further, in Table 2, we see that the average scores on a 5th grade state wide standardized test were 1308 points for math and 1246 points for reading and that there was an average of 0.877 recorded disciplinary offenses per student per school year.

[10]Throughout our analysis, a designation of subsidized lunch means that the student was tagged as receiving subsidized lunch at some point in the district's data from school years 1999 to 2009. Thus, it is a constant variable by student (as are race and sex).

Table 1: Descriptive Statistics, Part 1

		Initial Sample		Regular Lunch Sample		Subsidized Lunch Sample	
	District (K-12)	Tested	Admitted	Tested	Admitted	Tested	Admitted
Male	0.506	0.470	0.493	0.512	0.519	0.450	0.470
	(0.500)	(0.499)	(0.500)	(0.500)	(0.501)	(0.498)	(0.500)
African	0.559	0.319	0.169	0.119	0.070	0.490	0.312
American	(0.497)	(0.466)	(0.375)	(0.324)	(0.256)	(0.500)	(0.464)
Subsidized	0.772	0.569	0.429				
Lunch	(0.419)	(0.495)	(0.495)				
Income	28868	34341	38281	40677	43528	29139	31019
	(11153)	(13967)	(16058)	(15306)	(17105)	(9571)	(10626)
College	0.185	0.270	0.348	0.369	0.437	0.183	0.220
	(0.157)	(0.215)	(0.241)	(0.237)	(0.237)	(0.139)	(0.175)
Count	47506	1389	621	504	285	673	215

Standard deviations in parentheses.

Table 2: Descriptive Statistics, Part 2

		Initial Sample		Regular Lunch Sample		Subsidized Lunch Sample	
	District (K-12)	Tested	Admitted	Tested	Admitted	Tested	Admitted
Math	1308	1526	1610	1554	1624	1507	1598
	(221)	(177)	(182)	(185)	(198)	(163)	(161)
Reading	1246	1477	1547	1516	1560	1453	1533
	(217)	(163)	(150)	(160)	(135)	(157)	(151)
Offenses	0.877	0.189	0.066	0.073	0.027	0.245	0.088
	(1.884)	(0.950)	(0.521)	(0.376)	(0.183)	(0.910)	(0.460)
FSIQ		114.7	124.6	119.3	126.7	110.2	121.1
		(12.2)	(7.7)	(11.5)	(7.3)	(10.6)	(6.4)
1 year		0.891	0.918	0.891	0.923	0.899	0.921
retention		(0.311)	(0.275)	(0.312)	(0.267)	(0.302)	(0.270)
2 year		0.796	0.823	0.800	0.846	0.811	0.833
retention		(0.403)	(0.382)	(0.401)	(0.362)	(0.392)	(0.374)
Count	47506	1389	621	504	285	673	215

Standard deviations in parentheses.

To look specifically at the gifted program, we begin with a sample of 1389 students who were first tested for the gifted program in school years 2004-2005 to 2007-2008, were tested by a district psychologist, were attending a district school in the year they were tested, and were not tested multiple times for the gifted program.[11] Column 2 of Tables 1 and 2 report the characteristics of these students and Column 3 reports characteristics of those in this sample who were admitted into the gifted program.

We see that only 32 % of tested students are African American, compared to 56 % of district students, and the percent admitted is even lower at 17 %. Similarly, the proportion of subsidized lunch students that is tested is lower than the district proportion, and the proportion admitted is lower still. On average, compared to the district, tested students come from neighborhoods with a higher median income and a larger percentage of adults with at least a bachelor's degree, they score higher on a 5th grade standardized test in both math and reading, and they have fewer offenses. Admitted students come from even richer and more educated neighborhoods, score even higher on the standardized tests, and have fewer offenses.

From this sample, we exclude 72 students who were given a nonverbal test [12]

[11]We restrict our attention to students tested by district psychologists since a private psychologist hired by a student's parent may have incentives to inflate scores and scores would likely only be reported to the district if the score is above the admission cutoffs. Additionally, parents who hire a private psychologist may differ on unobservables. We focus on students who were attending a district school when tested in order to isolate retention effects of the program. While attraction effects are also of interest, they are beyond the scope of this paper. Finally, we do not consider students who were observed to have more than one set of test results for gifted consideration reported at any point up until summer 2009. Dropping observations with multiple tests removes potential manipulation of the type uncovered by Calcagno and Long (2008).

[12]Recall that nonverbal tests are given to students who are culturally or linguistically diverse. Thus, these students likely differ from the rest of the sample on unobservables.

and 1 student who took the test in 12th grade and therefore has no future retention outcomes available. Next we drop 138 students administered a test instrument other than the WISC4 [13] who consequently do not have the necessary sub-scores available for our modified RD analysis. Finally we drop one student who did take the WISC4 but does not have the necessary sub-scores available. Thus we we have a sample of 1177 students with 504 receiving regular lunch and 673 receiving subsidized lunch. Columns 4 and 5 of Tables 1 and 2 report the characteristics of the regular lunch sample and Columns 6 and 7 report the characteristics of the subsidized lunch sample. Since the admission criteria differ for students who receive subsidized lunch, we separately analyze these samples.

At the bottom of Table 2 we see that for the main sample, the unconditional mean of one and two year retention for tested students is 0.027 lower than that for admitted students. For the regular lunch sample the difference is .032 for one year retention and .046 for two year retention. For the subsidized lunch sample the differences are .021 and .022 for one and two year retention respectively. Thus, admittance into the gifted program does have some unconditional positive impact on retention.

3 Identification and Estimation

The starting point of our approach to identification and estimation is the fuzzy regression discontinuity (FRD) design. We discuss the consequences of manipulation of IQ scores and propose a modified estimator for the FRD design that deals with the potential manipulation of the forcing variable.

Following Fisher (1935), we adopt standard notation in the program evaluation

[13] 104 students took the WISC3 before the district had fully phased in the WISC4, 2 took the WJ-III, and 32 took the WPPSI

literature and consider a model with two potential outcomes.[14] Let D denote an indicator variable that is equal to one if a person receives treatment and zero otherwise. In this case, treatment is participation in the gifted program. Let Y_1 denote the outcome with treatment and Y_0 the outcome without treatment, where the outcome of interest is retention in the district. The researcher observes:

$$Y \;=\; D\,Y_1 + (1-D)Y_0 \tag{1}$$

The gain from receiving the treatment is defined as

$$\Delta \;=\; Y_1 \;-\; Y_0 \tag{2}$$

and note that this gain is unobserved for every single person in the sample. In terms of the treatment effect, the model can be written as

$$Y \;=\; Y_0 \;+\; D\,\Delta \tag{3}$$

The defining characteristic of the regression discontinuity design is that the treatment variables change discontinuously as a function of one or more variables. Here we focus on the "fuzzy" design" in which the probability of receiving the treatment changes discontinuously at certain points in the support of a forcing variable.[15] In our application the forcing variable is ability measured by an IQ score. Let Z denote the observed IQ score. Under the fuzzy design D is a random variable given Z. The propensity score defined as:

$$E[D|Z] \;=\; Pr\{D=1|Z\} \tag{4}$$

[14]This approach shares many similarities with the "switching regression" model introduced into economics by Quandt (1972), Heckman (1978, 1979) and Lee (1979). Heckman and Robb (1985) and Bjorklund and Moffitt (1987) treated heterogeneity in treatment as random coefficients model. It is also known in the statistical literature as the Rubin Model developed in Rubin (1974, 1978). See also Heckman and Vytlacil (2007) for an overview of the program evaluation literature.

[15]The "sharp" design is just a special case of the "fuzzy" design in which the probability of participation is either zero or one.

is known to be discontinuous at Z_0.

Following Hahn, Todd, and van der Klaauw (2001), identification of the treatment effect can be established using the following argument. Assume, for simplicity, that the treatment effect is constant. Let $e > 0$ denote an arbitrarily small positive number. Then

$$\begin{aligned} & E[Y|Z = Z_0 + e] \; - \; E[Y|Z = Z_0 - e] \\ = \quad & \Delta \left(E[D|Z = Z_0 + e] \; - \; E[D|Z = Z_0 - e] \right) \\ & + E[Y_0|Z = Z_0 + e] \; - \; E[Y_0|Z = Z_0 - e] \end{aligned} \tag{5}$$

The key assumption is then that $E[Y_0|Z]$ is continuous at Z_0. In that case, we have

$$\lim_{e \to 0} \; E[Y_0|Z = Z_0 + e] \; - \; E[Y_0|Z = Z_0 - e] \; = \; 0 \tag{6}$$

As a consequence, the treatment effect is identified from the following equation:

$$\Delta \quad = \quad \lim_{e \to 0+} \frac{E[Y|Z = Z_0 + e] \; - \; E[Y|Z = Z_0 - e]}{E[D|Z = Z_0 + e] \; - \; E[D|Z = Z_0 - e]} \tag{7}$$

Note that the denominator is nonzero because the fuzzy design guarantees that the propensity score is discontinuous at $Z = Z_0$.

The treatment effect can therefore be estimated as the ratio of two differences. Imbens (2007) discusses a simple computational approach for implementing this estimator that is based on local linear regression techniques. Using a uniform kernel with the same bandwidth for both the treatment and the outcome equation, he shows that one can characterize the estimator for Δ as a two stage least squares estimator. Using this approach, we can write the outcome equation as:

$$Y_i \quad = \quad \beta_0 + \beta_1 1\{Z_i < Z_0\}(Z_i - Z_0) + \beta_2 1\{Z_i \geq Z_0\}(Z_i - Z_0) + \Delta D_i + v_i \tag{8}$$

where Z_i is the observed IQ score.

Program participation is endogenous in the sense that $E[v_i|Z_i, D_i] \neq 0$. The selection equation for program participation is given by:

$$D_i = \alpha_0 + \alpha_1 1\{Z_i < Z_0\}(Z_i - Z_0) + \alpha_2 1\{Z_i \geq Z_0\}(Z_i - Z_0) + \alpha_3 1\{Z_i \geq Z_0\} + u_i \quad (9)$$

The key identification assumption in the RD design is then that $1\{Z_i \geq Z_0\}$ is a valid instrument for D_i.

Now if the IQ score is manipulated it is easy to see why the RD estimator fails. Consider the case in which households with positive shocks $v_i >> 0$ are more likely to manipulate the test score to make sure that their children get a test score above the threshold. In that case $E[v_i\ 1\{Z_i \geq Z_0\}] \neq 0$. As a consequences, the key identifying assumption of the RD estimator is not valid.

We argue that with additional information, we can develop a solution to address the manipulation of the forcing variable. In our case, we have access to additional intelligence measures, denoted by X_i. These measures consist of a set of sub-scores that are reported along will the full scale IQ score. While we find evidence for manipulation for the overall IQ score, we find that a set of the sub-scores which are not directly referenced in admission guidelines are not manipulated. The basic idea is to use these additional measures to predict an IQ score that is free of manipulation. Since parents are less likely to focus on sub-scores that are not directly relevant for admission, it may be reasonable to assume that such sub-scores are orthogonal to the error in the outcome equation.

To formalize these ideas, let $\hat{Z}_i$ denote the IQ score predicted by X_i, i.e. $\hat{Z}_i$ is a consistent estimator of $E[Z_i|\ X_i]$. Moreover, let us assume that the modified propensity score defined as:

$$E[D|\ \hat{Z}] \quad = \quad Pr\{D = 1|\hat{Z}\} \quad (10)$$

is discontinuous at Z_0. Note that this assumption needs to be tested. A discontinuity in $Pr\{D = 1|Z\}$ at Z_0 does not necessarily imply that $Pr\{D = 1|\hat{Z}\}$ also

has a discontinuity at Z_0. But in some cases the discontinuity will persist. These cases give rise to our modified RD estimator and the treatment effect is identified from the following equation:

$$\Delta = \lim_{e \to 0+} \frac{E[Y|\hat{Z} = Z_0 + e] - E[Y|\hat{Z} = Z_0 - e]}{E[D|\hat{Z} = Z_0 + e] - E[D|\hat{Z} = Z_0 - e]} \tag{11}$$

Again, we can implement this estimator using a 2SLS estimator. We only need to replace the observed IQ score with the predicted IQ score as the main forcing variable.

Summarizing the discussion, we have shown how to extend the RD estimator to allow for manipulation in the key forcing variable. Manipulation arises in our application because parents and administrators may have incentives to manipulate the test scores of students to help them obtain access to the gifted program. If there are additional sub-scores which are not directly referenced in the admission criteria, it is plausible to assume that these additional scores are uncorrelated with the error term in the main outcome equation. These sub-scores are correlated with the comprehensive IQ measures used to determine access to the program. Hence we can use the additional scores to predict IQ. We then use the exogenous part of the discontinuous change in program participation at the threshold that is predicted by the sub-scores as an instrument for program participation.

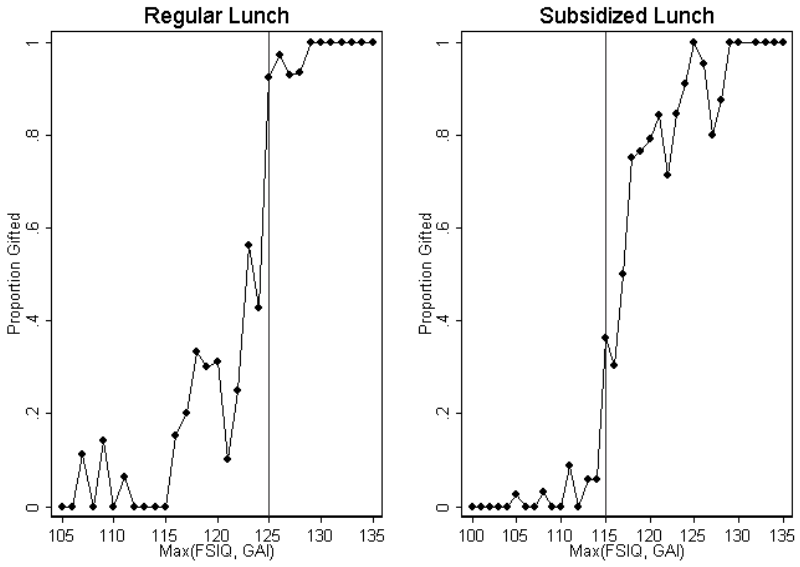
4 Empirical Results

4.1 IQ Testing and Admission

Since the district's regulations are in terms of both FSIQ and GAI, a natural starting point for the analysis is to consider the maximum of these two scores as the forcing variable that determines access to the gifted program. Figure 1 shows,

for the Regular and Subsidized Lunch samples, the proportion of students who are gifted as a function of the maximum of the FSIQ and GAI score.[16]

Figure 1: Proportion Gifted by Max(FSIQ,GAI)



Although there is some variability in both sample, we find that higher scores generally correspond to a higher proportion of students admitted into the gifted program. The cutoff of 125 for the regular lunch students and 115 for the subsidized lunch students is where there is the largest jump in proportion gifted, as expected.[17] Both our samples, therefore, meet the requirement that the probability of receiving treatment changes discontinuously at certain points in the support of the forcing variable (max(FSIQ,GAI)).

For a traditional fuzzy RDD approach to be valid, the distributions of any

[16]In this and subsequent figures, the running variable is truncated in order to focus on the area around the cutoff.

[17]Note that there are some students who are admitted into the program without meeting the IQ requirements. This is consistent with the requirement that students not be rejected based solely on not meeting the IQ thresholds.

covariates, including the forcing variable, should not show a discontinuity at the cutoff. Here we encounter a puzzling feature of the distribution of the main forcing variable. The maximum of FSIQ and GAI does not exhibit a smooth frequency distribution, especially for the Regular Lunch Sample. Figure 2 plots the two distribution that are heavily skewed to the right around the cut-off points for the two samples.[18] These spikes are robust to a number of sensitivity checks. For example, the patterns are not driven by any one administering psychologist or by testing in one particular grade or school year.[19] We view the above distribution as strong evidence of a concurrent discontinuity which potentially invalidates the use of the standard RD estimator as discussed in the previous section.[20]

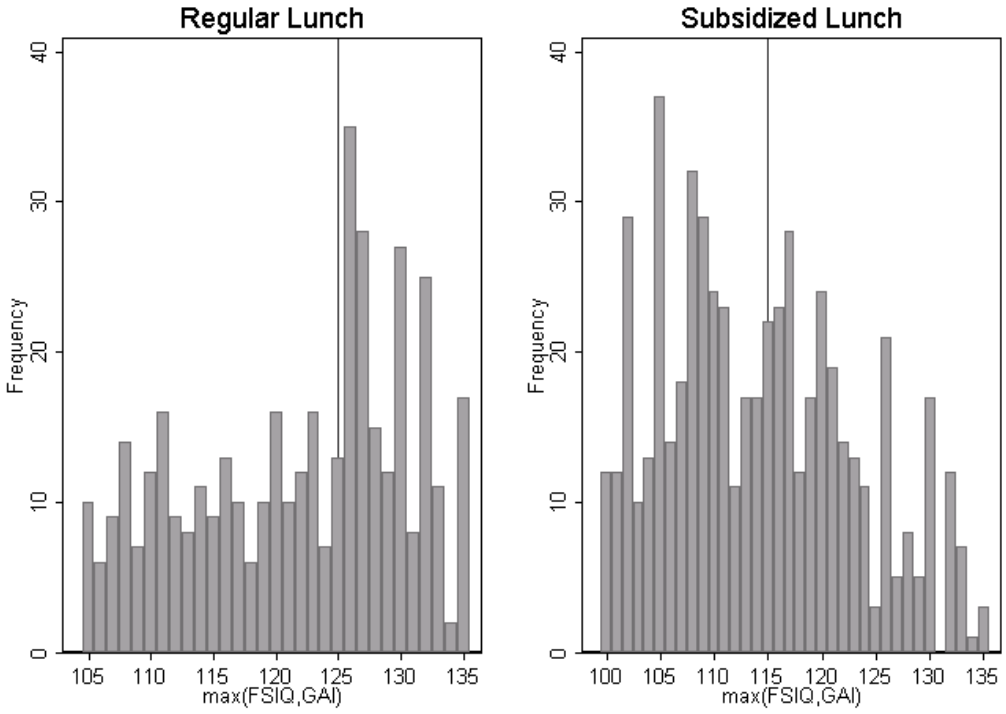
We, therefore, collected additional information to address this manipulation of the forcing variable. Specifically, in our samples, we have each student's Processing Speed Index (PSI) and Working Memory Index (WMI) scores along with their FSIQ and GAI scores. Recall that the PSI and WMI are not directly referenced in the admission guidelines and they are not used to calculate the GAI score. Table 3 reports descriptive statistics for the composite and index scores for the regular and subsidized lunch samples. For both samples, the PSI and WMI have

[18]The distributions are similarly spiked for FSIQ and GAI scores alone.

[19]Instead of relying purely on graphical evidence, one can also use a battery of formal tests. McCrary (2007), for example, has formalized the graphical procedure discussed above. He provides a framework for testing the null hypothesis of continuity of the underlying density function at the program cut-off points. In our application the visual evidence of manipulation in Figure 2 is sufficiently strong that that we clearly must address the likelihood that the running variable is manipulated.

[20]Martorell and McFarlin (2008) use RD to explore the effects of academic remediation which is required if a student scores below some cutoff value on a subject test. They find that the RD approach is valid when the score is based on a multiple choice test graded by a computer but not for a score based on a test graded by hand where the graders knew the necessary cutoff value. Our application is most similar to the latter case.

Figure 2: Score Distribution



the lowest averages among the scores. For the regular lunch sample, the PSI and WMI are the only scores positively skewed.

Table 3: Descriptive Statistics: Composites and Index Scores

Regular Lunch Sample

	Mean	Std. Dev.	Min	Max	Skew
max(FSIQ, GAI)	122.51	12.39	81	159	-0.240
FSIQ	119.28	11.53	74	149	-0.385
GAI	122.39	12.70	81	159	-0.227
VCI	119.02	12.64	89	152	-0.105
PRI	117.98	13.86	73	185	-0.115
PSI	107.90	12.73	68	147	0.076
WMI	111.55	13.64	59	148	0.022

Subsidized Lunch Sample

	Mean	Std. Dev.	Min	Max	Skew
max(FSIQ, GAI)	111.95	11.30	81	146	0.164
FSIQ	110.18	10.61	81	138	-0.020
GAI	111.49	12.19	74	146	0.093
VCI	108.60	11.61	69	144	0.439
PRI	108.65	12.14	65	143	0.002
PSI	105.90	13.13	68	151	0.004
WMI	106.07	11.72	68	144	0.175

We utilize the PSI and WMI to predict a score to use as the forcing variable. We predict the maximum of FSIQ and GAI using a linear and a quadratic approximation with PSI and WMI. Estimates from the regressions for these models are reported in Table 4. We find that PSI and WMI have predictive power in

explaining the maximum of FSIQ and GAI.

Table 4: Predicted Test Scores

	Regular Lunch		Subsidized Lunch	
	Linear	Quadratic	Linear	Quadratic
PSI	0.198***	0.522	0.236***	-0.489
	(0.037)	(0.539)	(0.026)	(0.369)
PSI^2		-0.00103		0.00182
		(0.002)		(0.001)
WMI	0.481***	0.885**	0.456***	-0.0952
	(0.032)	(0.450)	(0.030)	(0.405)
WMI^2		-0.00138		0.000988
		(0.001)		(0.002)
WMI*PSI		-0.000906		0.00321
		(0.003)		(0.003)
Constant	47.54***	7.957	38.57***	105.8***
	(5.014)	(44.300)	(4.108)	(33.520)
R-squared	0.362	0.364	0.345	0.35

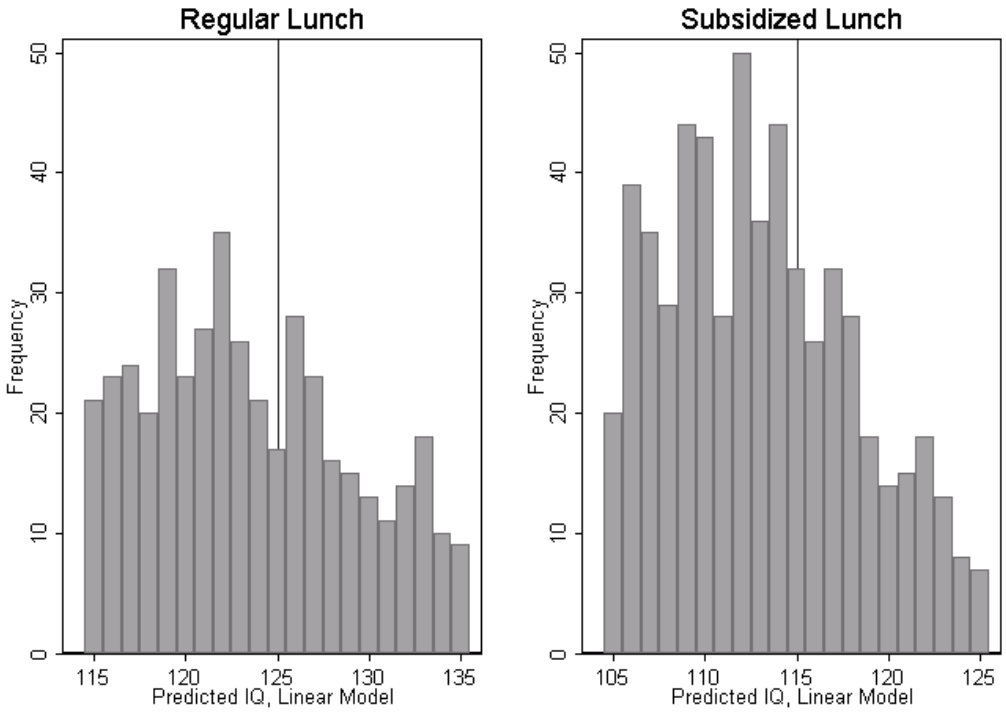
Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Figure 3 plots the distributions of the predicted scores from the linear model. We find that these distributions do not exhibit the anomalous behavior seen in Figure 2. The distribution for the Regular Lunch Sample is much smoother around the cutoff of 125 points and the distribution for the Subsidized Lunch Sample is smoother in general. Any discontinuities that remain are due to chance since there is no evidence of manipulation of the PSI and WMI scores. In the next section,

we show that there is still a cutoff value at which the probability of admittance changes discontinuously. Figure 4 shows the proportion of students gifted by predicted IQ from the linear model. We see that there is still an increasing probability of being gifted as the score increases.

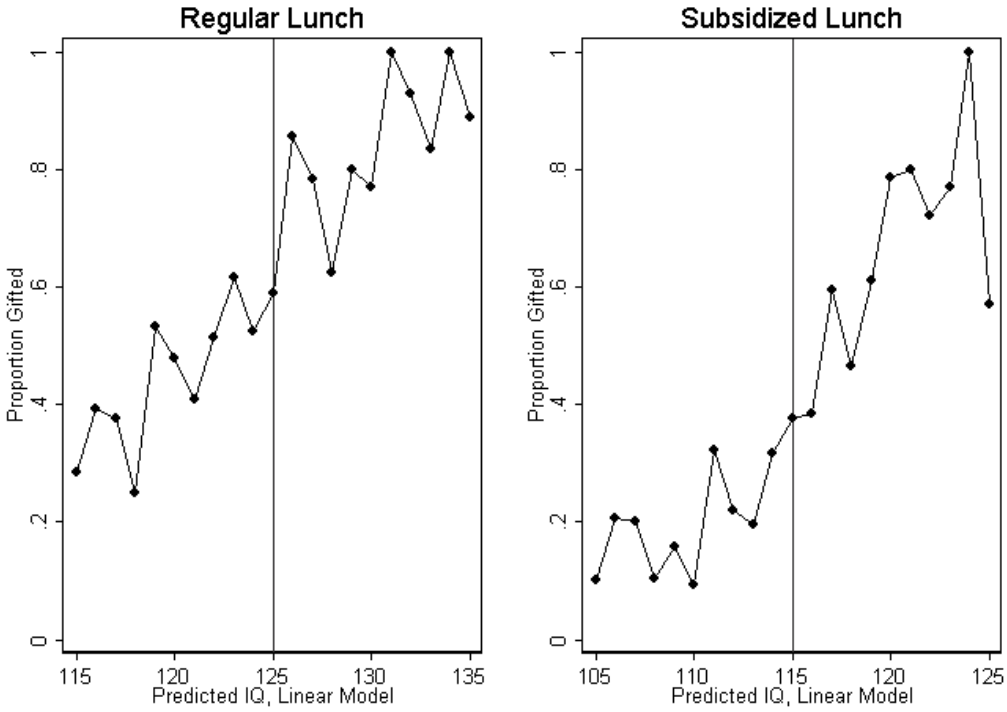
Figure 3: Distribution of Predicted IQ's from Linear Model



4.2 First Stage Estimates

Tables 5 and 6 present the results from the first stage of the RD estimator for the regular and subsidized lunch samples, respectively. For the first row of these tables the forcing variable is the maximum of FSIQ and GAI, for the second row the forcing variable is the predicted IQ from the linear model, and for the third row the forcing variable is the predicted IQ from the quadratic model. For each forcing

Figure 4: Proportion Gifted by Predicted IQ from Linear Model



variable, we report the number of observations within the given bandwidth[21], the difference between the right-hand side limit and the left-hand side limit of the conditional expectation of the probability of being gifted at the cut-off point, and the standard error of the difference. We find the difference using both local linear and local constant regressions. We consider the cutoff value of 125 points for the Regular Lunch Sample and 115 points for the Subsidized Lunch Sample. Note that the difference measures the degree of the discontinuity at the cut-off point. To make sure that our results are robust, we implement the estimators for different bandwidth choices.

Using the maximum of FSIQ and GAI as the forcing variable, we see that in the Regular Lunch Sample the estimates of the discontinuity at the cutoff of 125 points are positive, very large (about 0.6), and significant when we use the local constant estimator and smaller (0.2 to 0.4) but still positive and significant when we use the local linear estimator. [22]

Similarly, for the Subsidized Lunch Sample at the cutoff of 115 points, the estimates of the discontinuity are also positive, large(0.4 to 0.5), and significant with the local constant estimator and smaller (0.21 to 0.24) but still positive and significant with the local linear estimator.

[21]We include the cut-off point when we estimate the limit from above. Therefore, for a bandwidth of 4 points around the cutoff value of 125, the left hand side will include students with scores 121 to 124 (inclusive) and the right-hand side will include students with scores 125 to 128 (inclusive).

[22]Note that the forcing variable is measured in discrete increments of one which may partially explain the variability in the results using local linear regressions.

Table 5: Regular Lunch Sample: RDD First Stage, By Different Forcing Variables

		Bandwidth 4		Bandwidth 5		Bandwidth 6	
		Constant	Linear	Constant	Linear	Constant	Linear
Max(FSIQ, GAI)	Obs.	136	136	164	164	201	201
	Difference	0.589***	0.209***	0.607***	0.371***	0.624***	0.415***
		(0.076)	(0.029)	(0.065)	(0.020)	(0.059)	(0.017)
Linear Prediction Model	Obs.	171	171	208	208	254	254
	Difference	0.201***	0.019	0.212***	0.068***	0.212***	0.134***
		(0.073)	(0.018)	(0.066)	(0.014)	(0.060)	(0.011)
Quadratic Prediction Model	Obs.	176	176	218	218	262	262
	Difference	0.200***	0.018	0.231***	0.041***	0.220***	0.114***
		(0.072)	(0.018)	(0.065)	(0.014)	(0.059)	(0.011)

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Table 6: Subsidized Lunch Sample: RDD First Stage, By Different Forcing Variables

Subsidized Lunch Sample

		Bandwidth 4		Bandwidth 5		Bandwidth 6	
		Constant	Linear	Constant	Linear	Constant	Linear
Max(FSIQ, GAI)	Obs.	153	153	194	194	247	247
	Difference	0.388***	0.242***	0.457***	0.210***	0.522***	0.222***
		(0.061)	(0.013)	(0.054)	(0.009)	(0.047)	(0.007)
Linear Prediction Model	Obs.	241	241	312	312	369	369
	Difference	0.192***	-0.013	0.239 ***	-0.003	0.286 ***	-0.011
		(0.063)	(0.012)	(0.056)	(0.009)	(0.051)	(0.008)
Quadratic Prediction Model	Obs.	241	241	287	287	343	343
	Difference	0.197***	-0.051***	0.247***	-0.051***	0.280***	-0.003
		(0.066)	(0.015)	(0.061)	(0.012)	(0.056)	(0.010)

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Turning now to the second and third rows of Tables 5 and 6, we see that in both samples the first stage results are very similar between the two prediction models. In the Regular Lunch Sample, the estimates of the discontinuity are significant and positive (about 0.2) with the local constant estimator and smaller but still significant and positive (0.04 to 0.13) with the local linear estimator at bandwidths 5 and 6. In the Subsidized Lunch Sample, the estimates are significant and positive (0.19 to 0.29) with the local constant estimator. However, with the local linear estimator, the estimates of the discontinuity are small, negative, and mostly not significant.Thus, we have shown that for both samples and with each of the forcing variables, there is a significant discontinuity in the probability of being gifted at the cutoff score, at least in the local constant case.

4.3 Retention Effects

Next we investigate whether admittance into the gifted program helps the district retain students. We use one and two year retention (whether a student is in a district school one year or two years after being tested) as the outcome variables.[23] First, we implement OLS and IV estimators, with and without controls. Table 7 reports the estimated effects of gifted admittance on one and two year retention for both the Regular Lunch and Subsidized Lunch Samples. The instruments for the IV estimators are PSI and WMI.

For the Regular Lunch Sample, the OLS results show a positive and significant impact of being gifted ranging from 0.07 to 0.13. The IV results are also positive and significant and are slightly larger (0.12 to 0.15).

[23]A student is considered in the district two years after testing if s/he graduated from a district school one year after testing.

Table 7: Retention Estimates, Simple OLS and IV

	Regular Lunch Sample					Subsidized Lunch Sample			
	OLS		IV			OLS		IV	
	No Controls	Controls	No Controls	Controls		No Controls	Controls	No Controls	Controls
1 year retention					1 year retention				
Gifted	0.0735***	0.0811***	0.116*	0.129*	Gifted	0.0323	0.0324	0.125**	0.146**
	(0.028)	(0.031)	(0.060)	(0.071)		(0.025)	(0.027)	(0.059)	(0.065)
2 year retention					2 year retention				
Gifted	0.106***	0.132***	0.132*	0.150*	Gifted	0.0312	0.0523	0.114	0.157*
	(0.036)	(0.039)	(0.077)	(0.090)		(0.032)	(0.034)	(0.076)	(0.083)
Obs.	504	495	504	495	Obs.	673	661	673	661

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

For the Subsidized Lunch Sample, the OLS results indicate a positive but not significant impact of 0.03 to 0.05. The IV results are positive and mostly significant with impacts of 0.11 to 0.16. These results suggest that IQ scores are more likely to be manipulated for households that are more motivated to stay in the district.

Next, we present the RD estimates of the effect of gifted admission on retention. We focus on the results from the local constant specification since it gave the most consistent first stage results. As the forcing variable, we use the maximum of FSIQ and GAI and the predicated IQ's from the linear and quadratic prediction models. Recall that the density of the maximum of FSIQ and GAI shows a discontinuity at the cutoff and therefore violates the RD assumptions. We include it for comparison.

Tables 8 and 9 report the estimated effects for the Regular and Subsidized Lunch Samples, respectively. For the Regular Lunch Sample, using the maximum of FSIQ and GAI as the forcing variable, we find positive, though not significant, retention effects of 0.05 to 0.12. Using the predicted IQ's as the forcing variable gives positive and sometimes significant estimates of 0.18 to 0.41.

For the Subsidized Lunch Sample, using the maximum of FSIQ and GAI as the forcing variable, the retention effects are usually positive but never significant and range from -0.02 to .15. For Bandwidths 4 and 5, using the predicted IQ's as the forcing variable gives positive estimates of 0.1 to 0.3. Using the Bandwidth of 6, the estimates are smaller and sometimes negative, but not significant.

Thus, we find evidence that the gifted program does help the district retain students such as those in our samples, particularly in the case of regular lunch students. These findings suggest that the OLS and IV estimates reported in Table 7 may actually underestimate the effect of being admitted into the gifted program

for regular-lunch students, at least for those near the IQ cutoff. By contrast, we find overall smaller and relatively insignificant retention effects for the subsidized lunch students. This is to be expected. Relative to their wealthier counterparts, these poor households are less likely to have the resources to exit the district if their children fail to gain admission to the gifted program.

Table 8: Retention Estimates, Local Constant Regression, Regular Lunch Sample

	1 Year Retention				2 Year Retention		
	BW4	BW5	BW6		BW4	BW5	BW6
Max(FSIQ, GAI)							
Gifted	0.114	0.12	0.0697		0.0592	0.0461	0.0469
	(0.096)	(0.080)	(0.070)		(0.114)	(0.095)	(0.089)
Obs	136	164	201		136	164	201
Predicted IQ, Linear Model							
Gifted	0.406	0.328	0.332*		0.377	0.318	0.21
	(0.261)	(0.214)	(0.187)		(0.323)	(0.269)	(0.241)
Obs	171	208	254		171	208	254
Predicted IQ, Quadratic Model							
Gifted	0.35	0.221	0.292*		0.369	0.229	0.178
	(0.240)	(0.186)	(0.176)		(0.312)	(0.239)	(0.228)
Obs	176	218	262		176	218	262

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

BW = Bandwidth

Table 9: Retention Estimates, Local Constant Regression, Subsidized Lunch Sample

	1 Year Retention				2 Year Retention		
	BW4	BW5	BW6		BW4	BW5	BW6
Max(FSIQ, GAI)							
Gifted	-0.0227	0.0686	0.0069		-0.0227	0.149	0.0733
	(0.112)	(0.096)	(0.069)		(0.144)	(0.115)	(0.086)
Obs	153	194	247		153	194	247
Predicted IQ, Linear Model							
Gifted	0.118	0.101	-0.000997		0.214	0.148	0.0209
	(0.166)	(0.125)	(0.103)		(0.259)	(0.187)	(0.146)
Obs	241	312	369		241	312	369
Predicted IQ, Quadratic Model							
Gifted	0.279*	0.185	0.0279		0.299	0.238	0.0337
	(0.167)	(0.127)	(0.110)		(0.267)	(0.198)	(0.159)
Obs	241	287	343		241	287	343

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

BW = Bandwidth

5 Conclusions

We have studied the admissions procedures and retention effects of a gifted program in an urban school district. Gifted programs have the dual objective of engaging and challenging gifted students to reach advanced levels of achievement as well as attracting and retaining students who might otherwise leave for suburban or private schools. Many gifted programs, including the one studied in this paper, employ IQ thresholds for admission, with those above the threshold being admitted. However, a concern for transparency typically leads to promulgation of criteria for admission. Moreover, psychologists administering IQ tests may "give the benefit of the doubt" in assessing performance of students who are near the threshold for admission to a gifted program. These types of activities imply that the observed IQ scores are at best noisy measures of the underlying ability of the students. In many cases, the test scores appear to be manipulated. One consequence of this manipulation is that IQ scores at the cut-off point cannot be used as instruments for program participation, which invalidates the key identifying assumption of the standard regression discontinuity design.

We have shown in this paper how to extend the RD estimator allowing for manipulation in the key forcing variable. Manipulation arises in our application because parents and administrators have incentives to manipulate the IQ scores of students to help them obtain access to the gifted program. We have access to additional sub-scores that are not directly referenced in the admission guidelines. It is plausible to assume that these additional scores are uncorrelated with unobservable components in parental decisions to keep their children in district schools which is the main outcome variable of interest. We can use the additional test scores to predict the IQ. We then use the exogenous part of the discontinuous change in program participation at the threshold predicted by the sub-scores as an

instrument for program participation. Our findings suggest this approach works well in our application. Using the predicted scores yields plausible estimates of the propensity scores around the discontinuity. The results provide evidence that the program has a positive effect on retention for regular-lunch students. The district has recently revamped its gifted program and admissions procedures which may lead to less manipulation of IQ scores and even more positive retention effects in the future.

We view the results reported in this paper as promising for future research. Given the wide variety of potential applications of the regression discontinuity design in educational settings and the prevalence of potential manipulation of the forcing variable, the methods discussed in this paper may be of substantial importance for future research.

References

Angrist, J. and Lavy, V. (1999). Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement. *Quarterly Journal of Economics*, *114*, 533–75.

Argys, L. M., Rees, D. I., and Brewer, D. J. (1996). Detracking America's schools: Equity at zero cost?. *Journal of Policy Analysis and Management*, *15 (4)*, 623–645.

Betts, J. R. and Shkolnik, J. L. (2000). Key difficulties in identifying the effects of ability grouping on student achievement. *Economics of Education Review*, *19 (1)*, 21–26.

Bhatt, R. (2009). The Impacts of Gifted and Talented Education. Working Paper.

Bjorklund, A. and Moffitt, R. (1987). The Estimation of Wage and Welfare Gains in Self-Selection Models. *Review of Economics and Statistics*, *69*, 42–49.

Braddock, J. H. and Dawkins, M. (1993). Ability grouping, aspirations, and achievement: evidence from the National Longitudinal Study of 1988. *Journal of Negro Education*, *62 (3)*, 324–336.

Calcagno, J. and Long, B. (2008). The Impact of Postsecondary Remediation Using a Regression Discontinuity Approach: Addressing Endogenous Sorting and Noncompliance. CCRC Working Paper.

Chay, K., McEwan, P., and Urquiola, M. (2005). The Central Role of Noise in Evaluating Interventions that USE Test Scores to Rank Schools. *American Economic Review*, *95*, 1237–58.

Duflo, E., Dupas, P., and Kremer, M. (2008). Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya. NBER Working Paper 14475.

Epple, D., Newlon, E., and Romano, R. (2002). Ability Tracking, School Competition, and the Distribution of Educational Benefits. *Journal of Public Economics*, *83(1)*, 1–48.

Figlio, D. N. and Page, M. (20002). School choice and the distributional effects of ability tracking: does separation increase inequality?. *Journal of Urban Economics*, *51*, 497–514.

Fisher, R. (1935). *Design of Experiments*. Hafner. New York.

Gorman, A. (1987). The stratification of high school learning opportunities. *The Sociology of Education*, *60*, 135–155.

Hahn, J., Todd, P., and van der Klaauw, W. (2001). Identification and Estimation of Treatment Effects With Regression Discontinuity Design. *Econometrica*, *69 (1)*, 201–09.

Heckman, J. (1978). Dummy Endogenous Variables in a Simultaneous Equation System. *Econometrica*, *46*, 931–960.

Heckman, J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47 (1)*, 153–161.

Heckman, J. and Robb, R. (1985). Alternative Methods for Evaluating the Impact of Interventions. In *Longitudinal Analysis of Labor Market Data*. Cambridge University Press.

Heckman, J. and Vytlacil, E. (2007). Econometric Evaluation of Social Programs. In *Handbook of Econometrics 6b*. Elsevier North Holland.

Hoffer, T. B. (1992). Middle school ability grouping and student achievement in science and mathematics. *Educational Evaluation and Policy Analysis*, *14 (3)*, 205–227.

Imbens, G. (2007). Regression Discontinuity Design. NBER Lecture Notes: What's New in Econometrics?

Jacob, B. and Lefgren, L. (2004). The Impact of Teacher Training on Student Achievement: Quasi Experimental Evidence from School Reform Efforts in Chicago. *Journal of Human Resources*, *39*, 50–79.

Kerckhoff, A. (1986). Effects of Ability Grouping in British Secondary Schools. *American Sociological Review*, *51 (6)*, 842–858.

Lee, L. (1979). Identification and Estimation in Binary Choice Models with Limited (Censored) Dependent Variables. *Econometrica*, *47*, 977–996.

Martorell, P. and McFarlin, I. (2008). Help or Hindrance? The Effects of College Remediation on Academic and Labor Market Outcomes. Working Paper.

McCrary, J. (2007). Testing for Manipulation of the Running Variable in the Regression Discontinuity Design. *Journal of Econometrics*, *forthcoming*.

Oakes, J. (1990). *Multiplying inequalities: the effects of race, social class, and tracking on opportunities to learn mathematics and science*. RAND Corporation, Santa Monica, CA.

Quandt, R. (1972). A New Approach to Estimating Switching Regression Models. *Journal of the American Statistical Association*, *67*, 306–310.

Rubin, D. (1974). Estimating Causal Effects of Treatment in Randomized and Nonrandomized Studies. *Journal of Educational Psychology*, *66*, 688–701.

Rubin, D. (1978). Bayesian Inference For Causal Effects. *Annals of Statistics*, *6*, 34–58.

Thistlethwaite, D. and D.Cambell (1960). Regression-discontinuity Analysis: An Alternative to the Ex Post Factor Experiment. *Journal of Educational Psychology*, *51*, 309–17.

van der Klaauw, W. (2002). Estimating the Effect of Financial Aid Offers on College Enrollment: A Regression Discontinuity Approach. *International Economic Review*, *43 (4)*, 1249–87.

Vaughn, V. L., Feldhusen, J. F., and Asher, J. W. (1991). Meta-analysis and review of research on pull-out programs in gifted education. *Gifted Child Quarterly*, *35 (2)*, 92–98.

Zimmer, R. (2003). A New Twist in the Tracking Debate. *Economics of Education Review*, *22 (3)*, 307–315.